1999

Bootstrapping versus the student's t : the problems of Type I error and power

Laura L. Lansing
Lehigh University

Lansing, Laura L.

Bootstrapping versus the students's t; the problems of Type I error and power

May 31, 1999

Bootstrapping Versus the Student's t: The Problems of Type I Error and Power

by

Laura L. Lansing

A Thesis

Presented to the Graduate and Research Committee

of Lehigh University

in Candidacy for the Degree of

Master of Science

in

Psychology

Lehigh University

April 27, 1999

This thesis is accepted and approved in partial fulfillment of the requirements for the Master of Science in Psychology.

5/4/99
Date

Martin Richter
Thesis Adviser

Diane Hyland
Chairperson of Department

Acknowledgments

I am very grateful to Dr. Martin Richter for his patience and guidance during the research for and writing of this thesis, as well as for his ideas for how to adjust for inflated and deflated Type I error rates and how to maintain effect sizes, both of which proved to be very helpful. I am also grateful to Dr. Mark Bickhard and Dr. Gary Lutz for their participation as members of my thesis committee. I would like also to express appreciation to Dr. Edwin Kay for the time and effort he patiently gave in order to make modifications to the BNP program to make it more flexible. Finally, I would like to thank Mrs. Carol Richter, Dr. John Smith, and Dr. Kim Carroll-Smith for keeping me well fed during the entire degree process.

Table of Contents

List of Tables v

Abstract 1

Introduction 2

Method 18

Phase One: Type I 22

Phase Two: Power 30

Results and Discussion 32

Phase One: Type I 32

Phase Two: Power 36

General Discussion 40

Tables 46

References 62

Appendices 64

Appendix A: Confidence Intervals 64

Appendix B: Phase One Output 66

Appendix C: Unadjusted Phase Two Output 76

Appendix D: Adjusted Phase Two Output 85

Appendix E: Effect Size Output 94

Appendix F: Robust Rank Order Test 95

Appendix G: Program Implementation 97

Vita 98

List of Tables

Table 1: Various Sample Sizes and Ratios of Population Variances for Values chosen for Type I Error Rate Analysis 46

Table 2: Various Sample Sizes, Ratios of Population Variances for Values, and Means chosen for Power Analysis..... 49

Table 3: Actual Type I Error Rates when H_0 is True and Nominal Significance is $\alpha = .05$ for Student's t-test, Welch's t'-test, and the Pooled and Unpooled Bootstrap tests..... 52

Table 4: Actual Power when Nominal Significance is $\alpha = .05$ for Student's t-test, Welch's t'-test, and the Pooled and Unpooled Bootstrap tests (unadjusted α)..... 55

Table 5: Actual Power when Nominal Significance is $\alpha = .05$ for Student's t-test, Welch's t'-test, and the Pooled and Unpooled Bootstrap tests (ajdjusted α) 58

Table 6: Actual Power for Various Sample sizes when Effect Size is Maintained 61

Table 7: Actual Probability of Type I error for Sample sizes [2, 2] when H_0 is true and Nominal Significance is $\alpha = .05$ 67

Table 8: Actual Probability of Type I error for Sample sizes [5, 3] when H_0 is true and Nominal Significance is $\alpha = .05$ 68

Table 9: Actual Probability of Type I error for Sample sizes [5, 5] when H_0 is true and Nominal Significance is $\alpha = .05$ 69

Table 10: Actual Probability of Type I error for Sample sizes [20, 2] when H_0 is true and Nominal Significance is $\alpha = .05$ 70

Table 11: Actual Probability of Type I error for Sample sizes [50, 5] when H_0 is true and Nominal Significance is $\alpha = .05$ 71

Table 12: Actual Probability of Type I error for Sample sizes [50, 50] when H_0 is true and Nominal Significance is $\alpha = .05$ 72

Table 13: Actual Probability of Type I error for Sample sizes [100, 20] when H_0 is true and Nominal Significance is $\alpha = .05$ 73

Table 14: Actual Probability of Type I error for Sample sizes [200, 20] when H_0 is true and Nominal Significance is $\alpha = .05$ 74

Table 15: Actual Probability of Type I error for Sample sizes [200, 200] when H_0 is true and Nominal Significance is $\alpha = .05$ 75

Table 16: Actual Power for Sample sizes [2, 2] when Nominal Significance is $\alpha = .05$ 76

Table 17: Actual Power for Sample sizes [5, 3] when Nominal Significance is $\alpha = .05$ 77

Table 18: Actual Power for Sample sizes [5, 5] when Nominal Significance is $\alpha = .05$ 78

Table 19: Actual Power for Sample sizes [20, 2] when Nominal Significance is $\alpha = .05$ 79

Table 20: Actual Power for Sample sizes [50, 5] when Nominal Significance is $\alpha = .05$ 80

Table 21: Actual Power for Sample sizes [50, 50] when Nominal Significance is $\alpha = .05$ 81

Table 22: Actual Power for Sample sizes [100, 20] when Nominal Significance is $\alpha = .05$ 82

Table 23: Actual Power for Sample sizes [200, 20] when Nominal Significance is $\alpha = .05$ 83

Table 24: Actual Power for Sample sizes [200, 200] when Nominal Significance is $\alpha = .05$ 84

Table 25: Actual Power for Sample sizes [2, 2] when Actual Significance is $\alpha = .05$ for Student's t-test 85

Table 26: Actual Power for Sample sizes [5, 3] when Actual Significance is $\alpha = .05$ for Student's t-test 86

Table 27: Actual Power for Sample sizes [5, 5] when Actual Significance is $\alpha = .05$ for Student's t-test 87

Table 28: Actual Power for Sample sizes [20, 2] when Actual Significance is $\alpha = .05$ for Student's t-test 88

Table 29: Actual Power for Sample sizes [50, 5] when Actual Significance is $\alpha = .05$ for Student's t-test 89

Table 30: Actual Power for Sample sizes [50, 50] when Actual Significance is $\alpha = .05$ for Student's t-test 90

Table 31: Actual Power for Sample sizes [100, 20] when Actual Significance is $\alpha = .05$ for Student's t-test 91

Table 32: Actual Power for Sample sizes [200, 20] when Actual Significance is $\alpha = .05$ for Student's t-test 92

Table 33: Actual Power for Sample sizes [200, 200] when Actual Significance is $\alpha = .05$ for Student's t-test 93

Table 34: Actual Power for Various Sample sizes when Effect Size is Maintained 94

Abstract

Much psychological research is conducted using the analysis of variance in statistical analysis because it is easy to use and powerful. When the assumptions for the analysis of variance are not met, there are many nonparametric tests available. Unfortunately, nonparametric tests are not as powerful as the analysis of variance. A recent nonparametric alternative is the bootstrap statistic. The bootstrap is a resampling technique, which uses the distributional information in a sample while remaining distribution free. Additional advantages of the bootstrap are that it produces an estimate of how good an estimate it produces and that it can be used to study any statistic of interest.

In the present study, the Type I error rates and power of the bootstrap were explored. Using Monte Carlo simulations, two forms of the bootstrap statistic were compared to Student's t-test and Welch's t'-test. A bootstrap procedure with a pooled error term was compared to Student's t, and a bootstrap procedure with an unpooled error term was compared to Welch's t'. Three different power analyses were utilized in an effort to equate the bootstrap to the parametric tests. Results suggest that the bootstrap is just as powerful as the analysis of variance when sample sizes are large but does not perform well when sample sizes are small. Further research is needed to better understand the bootstrap. Additional work should make use of the power testing techniques utilized in the present study, namely, adjusting nominal significance to produce Type I error rates to the desired value and maintaining effect sizes across trials.

Bootstrapping Versus the Student's t: The Problems of Type I Error and Power

Many psychology experiments are analyzed using the analysis of variance (ANOVA). In its simplest form, and the form used in this investigation, the ANOVA is equivalent to the t or Student's t-test and can be used to make inferences about the difference between the means of two independent samples. Student's t statistic is as follows:

$$t = \frac{(\overline{Y_1} - \overline{Y_2}) - (\mu_1 - \mu_2)}{s_{\overline{Y_1} - \overline{Y_2}}}$$

where $\overline{Y_i}$ is the mean of the sample from population i, μ_i is the mean of population i, and $s_{\overline{Y_1} - \overline{Y_2}}$, the standard error of the difference between the means, is defined as $s_p(1/n_1 + 1/n_2)^{1/2}$ where s_p^2 is the pooled unbiased estimate of the population variance obtained from the two samples. The t-test has $(n_1 + n_2 - 2)$ degrees of freedom. In this way, the experimenter can test the null hypothesis of no difference between the population means, written H_0: $\mu_1 = \mu_2$, by creating a ratio of the between-group variability over the standard error of the difference between the means. Between-group variability is derived from the differences among sample means and is an estimate of the degree to which the population means differ plus experimental error. The standard error of the difference between the means, also referred to as the error term, is derived from an estimate of the population variances and assumes that both populations have the same variance. The standard error is an estimate of experimental error alone. When treatment effects are not present (i.e. H_0: $\mu_1 = \mu_2$ is true), the t-ratio is expected to be approximately equal to 1.00.

When treatment effects are present (i.e. H_0: $\mu_1 = \mu_2$ is false), the t-ratio is expected to be larger than 1.00. The experimenter's calculated t-ratio value is then compared to that in a standard table to determine whether the difference between the sample means is sufficiently large to suggest that the population means do in fact differ.

Researchers have good reason for their preference in using this test. Student's t-test is computationally simple relative to other methods. And, if the effects tested are in fact existent, one is likely to find them using a t-test with a sensitively designed experiment. In other words, Student's t-test methods are relatively easy to understand and conduct, and they are quite powerful. This is not to say that the t-test has no weakness. All statistical techniques are based on one or more assumptions, and Student's t is no exception. The assumptions for any analysis of variance are: the observed data are independent, the treatment populations from which the data are drawn are normally distributed, and the variances of these same populations are homogeneous. If these assumptions can not be met, Student's t-test may not be a suitable technique for analysis.

Under certain conditions, violation of an assumption upon which Student's t is based does not render the method inappropriate. Violation of the normality assumption has been shown to have little effect on an analysis of variance unless the populations from which the data are taken are highly skewed, the number of observations, n, is small (generally, $n_i <$ approximately 25 for all i where i = 1, 2, ..., a and a = the number of samples, depending on the application), or a unidirectional test is being employed (Kulkarni, 1993; Scheffe, 1959). Violation of the independence assumption has been shown to bias results, but the extent and direction of that bias is determined by the specific

form of dependence (Glass & Hopkins, 1996). Whether the bias is so great as to render Student's t-test inappropriate is up to the discretion of the analyst.

Moderate violation of the homogeneity of variance assumption has been shown to have little effect on a t-test when the same number of observations are in each sample ($n_1 = n_2 = \ldots = n_a$, where a = the number of samples). In cases where the homogeneity of variance assumption violation is great, there exists Welch's t'-test, which is a parametric t-test that makes an adjustment for the heterogeneity in the error term (i.e. the estimated standard error of the difference between the means) found in the denominator of the t-ratio (Bradley, 1993; Glass & Hopkins, 1996). Welch's t'-test is similar to Student's t in that it is a parametric approach to testing the difference between two independent sample means. The difference lies in the fact that Welch's t' makes a correction for violation of the homogeneity of variance assumption. The statistic is as follows:

$$t' = \frac{\overline{Y}_1 - \overline{Y}_2}{s_{Y_1 - Y_2}} = \frac{\overline{Y}_1 - \overline{Y}_2}{[(s_1^2/n_1) + (s_2^2/n_2)]^{1/2}}$$

where $\overline{Y}_i$ is the mean for the sample from population i, $s_{Y_1 - Y_2}$ is the sample estimate of the standard error of the difference between the means, and s_1^2 and s_2^2 are the sample variances from populations one and two, respectively. Welch's t' has the following degrees of freedom:

$$df' = \frac{(s_1^2/n_1) + (s_2^2/n_2)}{\frac{[s_1^2/n_1]^2}{n_1 - 1} + \frac{[s_2^2/n_2]^2}{n_2 - 1}}$$

The degrees of freedom for this test are based on Satterthwaite's approximation for a t-test when correcting for heterogeneity of variances (Bradley, 1993).

Simultaneous violation of both the normality and homogeneity of variance assumptions has been shown to affect Student's t in an additive fashion. In some cases the violations will counteract each other, and in other cases each will exaggerate the effect of the other (Glass & Hopkins, 1996).

One measure of how well a significance test is working under assumption violation is how well it controls the probability of Type I error, which is defined as the probability of incorrectly rejecting the null hypothesis. The desired value for the probability of a Type I error, that is the frequency with which an experimenter is willing to risk mistakenly rejecting the null hypothesis when in fact there are no treatment effects, is denoted by α_n for "nominal alpha." Nominal alpha is also known as the significance level of the test and is determined by the experimenter before data collection. Most often α_n is set at .05 or .01, meaning that approximately five times or one time, respectively, out of 100 an experimenter can expect to incorrectly reject the null hypothesis. The empirical value for the probability of a Type I error in a given experiment is denoted by α_a for "actual alpha." Actual alpha is the proportion of times the null hypothesis is rejected in an actual experiment when in fact there are no treatment effects present. This value is not determined by the experimenter; rather, it must either be found empirically or derived mathematically. When all assumptions are met $\alpha_a = \alpha_n$ within measurement error.

Simultaneous violation of the homogeneity of variance assumption and unequal sample sizes is known to bias the probability of a Type I error when conducting a

Student's t-test. Actual alpha for the t-test has been shown to exceed α_n when the larger samples are associated with the smaller variances (Hsu, 1938; Kulkarni, 1993). Similarly, α_a has been shown to be less than α_n when the larger samples are associated with the larger variances (Hsu, 1938; Kulkarni, 1993). In such situations, Welch's t'-test may be an acceptable alternative to Student's t.

Welch's t'-test produces Type I error rates closer to nominal than Student's t in cases of heterogeneous variances because of the adjustment it makes in the error term (i.e. the estimate of the standard error of the difference between the two means) and in the degrees of freedom used in the test. In testing for the difference between two means, Student's t assumes homogeneity of variances. Consequently, it uses a single value to estimate both population variances. Welch's t'-test, on the other hand, does not assume homogeneity of variances. Consequently, it uses two estimates for the population variances, one for each sample, and weights these proportionately by sample size in estimating the standard error. These two different types of estimation for the standard error of the difference between the means have sufficient impact on the tests that they have been given specific names. The error term used in Student's t is called a pooled error term, reflecting the use of all data from both groups to best calculate a single common variance estimate. The error term used in Welch's t' is called an unpooled error term, reflecting the use of a separate variance estimate for each group. The use of the unpooled error term means that Welch's t' results in only an approximation of the t distribution. Using the Satterthwaite calculation for the degrees of freedom when using a Welch's t' test makes the approximation better.

Another measure of how well a significance test is working under assumption violation is how high it makes the power of the test, which is defined as the probability of finding statistically significant effects if such effects do, in fact, exist. Power is denoted by $(1 - \beta)$, where β is called the probability of a Type II error which is defined as the probability of incorrectly retaining the null hypothesis. Power is typically used in conjunction with the probability of a Type I error to establish the degree of trust the experimenter has in a test. The power of a test may vary with the application, but generally a power of 0.70 or higher is considered acceptable in most psychological applications.

When assumption violations or bias in the probability of Type I or Type II errors render the t-test or t'-test inappropriate, there are a number of so called nonparametric techniques available to the experimenter. Nonparametric tests make fewer and weaker assumptions about the distribution of the data than do parametric statistics. And, nonparametric techniques tend not to use all of the information provided in a sample. For example, many nonparametric statistics compare sample medians. The calculation of a median is not sensitive to extreme data points in that only one or two data points that fall in the middle of the sample with respect to magnitude are directly utilized; and consequently, most of the information about the population distribution in the remaining data is lost. Student's t and Welch's t'-test, by contrast, compare sample means. The calculation of a mean is sensitive to extreme points in that all data points are equally weighted; and hence, the mean uses more information inherent in the data set. Generally speaking, nonparametric statistics are not as powerful as the parametric Student's t and

Welch's t'-tests (Siegel & Castellan, 1988). But, when one can not be sure of the methods used for data collection, when the data do not appear to be even remotely normally distributed, or when the variances are clearly not homogeneous, nonparametric techniques can make the difference between a slightly less powerful result and scrapping the entire experiment (Siegel & Castellan, 1988).

A relatively recent nonparametric technique available when an analysis of variance is inappropriate is the bootstrap. The name "bootstrap" comes from the folk saying "to pull ones self up by the bootstraps" and refers to the procedure's ability to provide an estimate of a population characteristic while simultaneously providing a measure of the precision, or error, in the estimate. By using a resampling technique, the bootstrap exploits all of the information in a single data set to make inferences about the unknown population (Diaconis & Efron, 1983; Efron & Tibshirani, 1993; Strube, 1988).

To make use of the bootstrap, a random sample of size n is taken from the population of interest. This original random sample is treated as if it were, in fact, the population. A large number of random samples of size n, called bootstrap samples, are taken with replacement from this initial sample. For each bootstrap sample, the statistic of interest is calculated and is called a bootstrap replication. The sampling distribution of the statistic is approximated by the distribution of these bootstrap replications. Any statistical test of interest can then be conducted using the estimated sampling distribution found through the bootstrapping procedure to make inferences about the population.

For example, in this study the statistic of interest was the difference between two sample means. Suppose the following two random samples are taken from two identical

populations. Let $Y_1 = \{-4, -2, 1, 3, 7\}$ be a sample of size n = 5 from one population with $\mu_1 = 0$ and $\sigma_1^2 = 25$. Let $Y_2 = \{1, 2, 3, 4, 4, 6, 7\}$ be a sample of size n = 7 from another population with $\mu_2 = 0$ and $\sigma_2^2 = 25$. The sample mean for the first sample is $\overline{Y_1} = 1.00$, and the sample mean for the second sample is $\overline{Y_2} = 3.86$. The difference between the population means is $\delta = 0 - 0 = 0$, whereas the difference between the sample means is $d = 1.00 - 3.86 = -2.86$. In the bootstrapping process, these random samples are treated as if they are, in fact, the populations from which they were taken.

In classical statistical theory, an observation can be expressed as $Y_{i,j} = \mu_i + \varepsilon_{i,j}$ where $Y_{i,j}$= the jth observation from sample i, μ_i= the mean of the population from which sample i was taken, and $\varepsilon_{i,j}$= the error present in the jth observation of sample i. As a population parameter, μ_i is fixed, and each observation in sample i can be thought of as varying from μ_i by some random amount $\varepsilon_{i,j}$. The value of an error score, $\varepsilon_{i,j}$, can be estimated by $Y_{i,j} - \hat{\mu}_i$, where $\hat{\mu}_i = \overline{Y}_i$, the mean of sample i and the best estimate available for the population mean. The standard ANOVA assumes normality which means that with knowledge of μ_i and σ_i^2 the population distribution can be determined exactly. The best estimate of μ_i is $\overline{Y}_i$, and the best estimate of σ_i^2 is the sample variance, s^2. The ANOVA also assumes homogeneity of variances which means that in calculating s^2 the experimenter can use the error scores from all samples to get the most accurate estimate for the population variance.

The bootstrap procedure, on the other hand, assumes only that the original samples are randomly drawn from their populations, suggesting that the samples are representative of those populations. Consequently, $\overline{Y}_i$ is the best estimate available for μ_i, and each

observation, $Y_{i,j}$, can be found by adding an error score to the best estimate of μ_i. Error scores can be estimated by $(Y_{i,j} - \hat{\mu}_i)$. If the experimenter further assumes homogeneity of variances, the error scores obtained from each sample can be pooled to create the most accurate estimate of the population variance.

In this example homogeneity of variances is assumed, and an error pool is created by calculating the differences between each observation and its sample mean. The error scores are:

Y_1 errors:	Y_2 errors:
-4.00 - 1.00 = -5.00	1.00 - 3.86 = -2.86
-2.00 - 1.00 = -3.00	2.00 - 3.86 = -1.86
1.00 - 1.00 = 0.00	3.00 - 3.86 = -0.86
3.00 - 1.00 = 2.00	4.00 - 3.86 = 0.14
7.00 - 1.00 = 6.00	4.00 - 3.86 = 0.14
	6.00 - 3.86 = 2.14
	7.00 - 3.86 = 3.14

Under the assumption of identical populations, the error pool is the set of the resultant $(n_1 + n_2) = 12$ error scores, {-5.00, -3.00, -2.86, -1.86, -0.86, 0.00, 0.14, 0.14, 2.00, 2.14, 3.14, 6.00 }. This error pool along with the means of the sample observations, $\frac{\sum_{j=1}^{n_1} Y_{1,j}}{n_1} = 1.00$ and $\frac{\sum_{j=1}^{n_2} Y_{2,j}}{n_2} = 3.86$ are used to create a large number, say B, of bootstrap replications.

A set of $n_1 = 5$ and a set of $n_2 = 7$ error scores are drawn with replacement from the pool of twelve error scores and their means calculated. One such drawing might result in the following two error means: [(-0.86) +0.00 + 2.14 + 6.00 + (-3.00)] /5 = 0.86 and

$$[2.00 + 2.00 + 3.14 + (-5.00) + (-2.86) + (-0.86) + 0.14]/7 = -0.21$$

These values are then used to estimate bootstrap means for the two populations, as well as their difference. The bootstrap means are $\overline{Y^*_{1,j}} = 1.00 + 0.86 = 1.86$ and $\overline{Y^*_{2,j}} = 3.86 + (-0.21) = 3.65$. A bootstrapping replication, d_j^*, is defined as the difference between the two jth (where j = 1, 2, ..., B) bootstrapping means, so $d_j^* = 1.86 - 3.65 = -1.79$. This bootstrapping procedure is repeated B times.

The resulting set of B bootstrap replications is used to create the sampling distribution of the estimate of the difference between the population means. The mean difference, the variability in the differences, or any other meaningful statistic of interest can be examined in order to make inferences about the difference between the population means. The best estimate of δ, the difference between the population means, is the difference between the original sample means, d.

The bootstrap technique was introduced by Efron (1979) as an alternative to the usual nonparametric methods. The technique falls between parametric methods such as Student's t and Welch's t' and standard nonparametric methods in that it avoids distributional assumptions and yet is not distribution free (Lunneborg, 1985; Strube, 1988). That is, one need not know anything about the population distribution to use the bootstrap, but the technique uses all the distributional information inherent in the sample to create an estimate of the sampling distribution of the statistic of interest. Consequently, the bootstrap provides the analyst with another option when the t-test assumptions about the population are unwarranted, but the loss of information incurred when using standard nonparametric tests is also undesirable (Strube, 1988).

Efron (1979) designed the bootstrap technique to use computational power as a means for obtaining an estimate of the standard error of any statistic, whether or not the statistic is mathematically tractable or the population distribution is known (Diaconis & Efron, 1983). One measure used to test the bootstrap's performance in the current study was the actual Type I error rate, or α_a. Type I errors were calculated by creating confidence intervals for the difference between the means and then counting the number of confidence intervals that did not include the null hypothesis of equal means. If zero was not included within an interval's limits, that interval did not include the null hypothesis and was included in the count. Dividing the result by the total number of bootstrap samples yielded an estimate of the actual Type I error rate. For each set of B bootstrap replications, a 95% confidence interval was constructed. The number of confidence intervals that did not include zero was then calculated. If C was the number of such intervals, then, $\alpha_a = C/B$.

Another measure used to test the bootstrap's performance in the current study was power, or the probability of finding a difference between the two means when such a difference was present. Typically, in practical applications, power can not be calculated directly because β is not known directly. If we can assume normal distributions and equal variances, power can be estimated as a function of n, α_n, and ω^2, an estimate of the effect size (if some reasonable estimate for ω^2 can be found), by using Cohen power tables or Pearson-Hartley charts. Alternatively, power can be found empirically through simulation studies, as was done in this investigation. Power was calculated by the same technique used in finding Type I error rates. For each set of B bootstrap replications, a 95%

confidence interval was constructed. If C was the number of such intervals which did not include zero, then $(1 - \beta) = C/B$. When investigating Type I error rates, data were simulated from populations with equal means. When investigating power, data were simulated from populations with unequal means.

This technique for calculating power works well as long as homogeneity of variances is maintained. Power calculation is affected by Type I error rates, and when all ANOVA assumptions are met, $\alpha_a = \alpha_n$ within measurement error. When homogeneity of variance is not maintained, actual Type I error rates are either inflated or deflated. In such cases, interpreting power is problematic when one uses the $(1 - \beta) = C/B$ formulation discussed above. Power values are generally cited with the understanding that had there been no difference between the means, α_a would have been equal to the nominal value of .05. What does it mean practically when a power of, say, 0.89 is obtained for a case in which the actual Type I error rate would have been 0.69 had the population means been equal? How can a comparison be made between such a case and one for which $\alpha_a = \alpha_n = .05$ when the population means are equal and a power of 0.75 is obtained when the population means are not equal?

In order to address this problem, a researcher can run Monty Carlo simulations to vary the value for nominal alpha until a value for α_n is found that produces $.04 \leq \alpha_a \leq .06$, a value sufficiently close to $\alpha_a = .05$. This makes it possible to compare the relative power for different cases because each case then produces a Type I error rate of .05 when the population means are equal. Under these conditions, it is then possible to use the $(1 - \beta) = C/B$ formulation discussed above to calculate power values with relative practical

meaning.

The simplest type of confidence interval, and the one used in the current study, is called the percentile method. This method assumes that the populations are symmetric about the median (and, consequently, the mean) although it does not assume that the data themselves are normally distributed (Efron, 1988; Efron & Gong, 1983; Kulkarni, 1993; Strube, 1988). A percentile method confidence interval for a bootstrap estimate is calculated as follows. The bootstrap replications, $d_j^* = (\overline{Y_1^*} - \overline{Y_2^*})$ where j = 1, 2, ..., B are listed in ascending order. The lower limit for the interval is then defined as the bootstrap estimate in rank$((\alpha/2)(B + 1))$, and the upper limit for the interval is defined as the bootstrap estimate in rank$((1-\alpha/2)(B + 1))$. When these calculations do not result in integral values, Buckland (1984) recommends linear interpolation or rounding to the nearest integer. When the data are not symmetric about the median (and hence, the mean), there are several methods available for obtaining adjusted bootstrap confidence intervals, all attempting to control for varying degrees of deviation from symmetry about the median in the underlying population. Without such a correction, deviation causes bias in percentile confidence intervals. A discussion of some of these corrective methods can be found in Appendix A.

While there appears to be agreement in the field regarding the bootstrap as a possible solution to some problems where a parametric solution is untenable (Efron, 1979; Efron & Gong, 1983; Efron and Tibshirani, 1993; Kulkarni, 1993; Strube, 1988), the bootstrap has been criticized for seriously inflated Type I error rates (Kulkarni, 1993; Rasmussen, 1987). When 500 bootstrap samples were drawn from initial random samples

of n = 5, 15, 30, and 60 data points in the estimation of correlation coefficients, the bootstrap had unacceptably high Type I error rates (0.171, 0.147, 0.139, and 0.120, respectively) compared to those of the standard parametric solutions (0.050, 0.052, 0.045, and 0.047, respectively) at the $\alpha_n = .05$ level (Rasmussen, 1987).

Kulkarni (1993) studied Type I error rates using two bootstrapping procedures to test the difference between two independent means. A bootstrap procedure using a pooled error term produced Type I error rates comparable to those of the t-test for large, unequal sample sizes and heterogeneous variances when using the percentile method for determining confidence intervals. An unpooled bootstrap procedure produced Type I error rates closer to $\alpha_n = .05$. The pooled bootstrap procedure produced Type I error rates of 0.329, 0.122, and 0.004 in cases where $n_1 = 100$, $n_2 = 20$ and variance ratios were 1:5, 1:2, and 5:1, respectively. For the same cases, the t-test produced Type I error rates of 0.331, 0.116, and 0.004, respectively. The bootstrap procedure using an unpooled error term produced Type I error rates closer to the nominal value of .05 (0.074, 0.070, and 0.055, respectively, for the same cases as above). Kulkarni (1993) found further that the pooled and unpooled bootstrap methods produced inflated Type I error rates with small sample sizes as well. The Type I error rate when $n_1 = 5$, $n_2 = 3$, and the variance ratio was 1:10 was .285 for the pooled bootstrap method and .243 for the unpooled bootstrap method. The Type I error rates were .131, .131, and .125 when $n_1 = n_2 = 5$ and the variance ratios were 1:10, 1:5, and 1:1, respectively, for the pooled bootstrap method and .141, .129, and .130, respectively, for the unpooled bootstrap method.

Researchers have yet to investigate bootstrapping power relative to the power of

standard parametric methods, though it has been suggested that the bootstrap should be less powerful than parametric and possibly some nonparametric techniques when the distribution is more normal than the sample data (Lunneborg & Tousignant, 1985).

The purpose of this study was to replicate and expand upon the work of Kulkarni (1993). Kulkarni examined the effects upon α_a of various degrees of heterogeneity of variance in conjunction with differing combinations of sample size when using two forms of the bootstrapping procedure, the parametric t-test, and the Robust Rank Order test, a nonparametric test used to test for equality of sample medians. The two bootstrapping procedures included one test that used a pooled error term and one that used an unpooled error term (Kulkarni, 1993) in the calculation of the standard error of the difference between the means. The procedure using the pooled error term assumes homogeneity of variance and thus was expected to result in Type I error rate biases similar to those found for the t-test, which also assumes homogeneity of variance. The procedure using the unpooled error term does not assume homogeneity of variance and so was expected to perform better than the t-test when the assumption of homogeneity of variance is violated. Kulkarni did not examine the case when sample sizes are unequal but homogeneity of variance is maintained, nor did she look at large equal sample sizes or cases where one sample is large (defined as $n \geq 20$ in this study) and the other small (defined as $n \leq 5$ in this study).

This study addressed these omissions as well as tried to validate further Kulkarni's (1993) results through replication. In addition to these tests and instead of the Robust Rank Order test the current study examined the performance of Welch's t'-test, a

parametric t-test corrected for heterogeneous variances (Bradley, 1993; Glass & Hopkins, 1996). The current study also examined the power of the bootstrap. The researcher expected to find that the parametric t-test results in the most accurate Type I error rates (i.e. α_a more nearly equal to α_n) when all Student's t-test assumptions are met. The researcher expected to find that when the homogeneity of variance assumption is violated, Welch's t'-test results in empirical Type I error rates most nearly equal to α_n because Welch's t' is a parametric test which adjusts for heterogeneity of variances. The researcher expected to find the pooled bootstrapping method superior (i.e. α_a more nearly equal to α_n) to the unpooled method when all Student's t-test assumptions are met because the pooled bootstrapping method uses an error term similar to that of Student's t, which assumes homogeneity of variances. The researcher expected to find the unpooled bootstrapping method superior to the pooled method when the homogeneity of variance assumption is violated because, like Welch's t', the unpooled bootstrapping method does not assume homogeneity of variances. Further, the researcher expected to find, following Kulkarni (1993) and Rasmussen (1987), that overall bootstrapping methods result in Type I error rates closest to α_n when sample sizes are large.

The researcher expected the power of the bootstrap to be inferior to that of Student's t-test when all t-test assumptions are met. Student's t assumes normality which means that the population distribution can be determined exactly if the population mean (μ) and variance (σ^2) are known. When μ and σ^2 are not known, $\bar{Y}$ and s^2, respectively, are the best estimates available, allowing the experimenter to get a good approximation of the population distribution. The ability to make the normality assumption is what gives

Student's t its power. The bootstrap, on the other hand, is a nonparametric test and requires only that the sample be randomly selected from the population. The fact that the bootstrap does not make any assumption about the population distribution suggests that the bootstrap can be expected to have lower power.

The researcher expected Welch's t'-test to be more (less) powerful than Student's t when the homogeneity of variance assumption was not met and the larger sample was paired with the larger (smaller) variance. The researcher expected the Phase One analysis to indicate that Student's t-test produces deflated (inflated) Type I error rates for such cases. This would mean that Student's t-test was producing 95% confidence intervals for the difference between the means that are too large (small). Consequently, this would result in lower power for Student's t. Since Welch's t'-test adjusts for the heterogeneity of variance, the researcher expected it to result in Type I error rates closer to the nominal $\alpha_n = .05$. The increase (decrease) in α_a was expected to result in an increase (decrease) in power as well.

The researcher expected the unpooled bootstrap to be more powerful than the t-test when Student's t-test assumptions are not met and sample sizes are large because the unpooled bootstrapping method attempts to adjust for heterogeneity of variance, where Student's t does not. Large sample sizes are stipulated here because bootstrapping methods assume the original sample is representative of the population, an assumption which is more likely to be met when sample sizes are large.

Method

This study consisted of two phases. The first phase involved the collection of

several pairs of data sets to be analyzed for Type I error rates. These data were generated using DATASIM (Bradley, 1993), a simulation software package designed specifically for statistical research. The data sets were normally distributed with means of zero and various degrees of heterogeneity of variance. The hypothesis that two data sets have identical means was tested in several ways at the $\alpha_n = .05$ level. Student's t and Welch's t' were implemented in DATASIM to conduct the test using parametric methods. Student's t also was implemented in a computer program called BNP (Kulkarni, 1993), as were a pooled and an unpooled bootstrap method. The latter two tests were used to conduct the test of equal means using bootstrapping techniques. The performance of each method was tested by creating a 95% confidence interval about the discrepancy between the true difference between the original population means (i.e. zero) and the empirical difference between the original sample means (d) using the bootstrap replications to estimate the standard error for each sample. If zero was not in the confidence interval, then one was added to the count of Type I errors. Each statistical test's performance was then measured by the extent to which α_a for a test deviated from $\alpha_n = .05$.

The second phase involved a similar collection of data sets to be used in analyses of power. These data also were generated using DATASIM (Bradley, 1993). The data sets were normally distributed, and the populations from which they were drawn had unequal means. Again, the data sets were sampled from populations with various degrees of heterogeneity of variance. The hypothesis that two data sets have identical means was tested in the same manner as was done in Phase One. Student's t and Welch's t' were implemented in DATASIM to conduct the test using parametric methods, and Student's t

was implemented again in the BNP program (Kulkarni, 1993) along with the pooled and unpooled bootstrap methods. The latter two tests were used to conduct the test of equal means using the bootstrapping technique. The performance of each method was tested by creating a 95% confidence interval about the empirical difference between the original sample means (d) using the bootstrap replications to estimate the standard error for each sample. If zero was not in the confidence interval, then one was added to the number of times the null hypothesis was correctly rejected. The proportion of correctly rejected tests was equal to the power of the test. The fact that the Type I error rates for the various cases were so different suggested that the power analysis may have been ambiguous.

Power is conceptualized as the probability of finding existent differences when $\alpha_a = .05$ (or in some cases .01) had the null hypothesis of equal means been true. To make meaningful comparisons, it was necessary to find α_n for which $\alpha_a = .05$ for all cases where Type I error rates were either deflated or inflated (i.e. $\alpha_a << .05$ or $\alpha_a >> .05$, respectively) due to combinations of unequal sample sizes and various degrees of heterogeneity of variance. In all but seven of these cases, this adjusted value was used as α_n in a second power analysis, producing a value for power which could then be compared with the other tests used in the study. In these cases, $(1.00 - \alpha_n)\ x\ 100\%$ confidence intervals were created about the empirical difference between the original sample means (d) using the bootstrap replications to estimate the standard error for each sample. Again, if zero was not in the confidence interval, then one was added to the number of times the null hypothesis was correctly rejected. The proportion of correctly rejected tests was equal to the power of the test, the same as with the previous analyses.

The remaining seven cases were not analyzed a second time with an adjusted α_n because of their extremely high Type I error rates in the first phase of the study. In each of these cases, $n_1 > n_2$ and $\sigma_1 < \sigma_2$, and the Type I error rates ranged from .226 to .433. In order to produce an actual Type I error rate of .05, the nominal alpha had to be adjusted to a value equal to or less than .001. In effect, this meant that 100% confidence intervals would need to be used in the power analysis.

This analysis utilized an adjusted α_n obtained for Student's t-test only. Adjusting the Student's t nominal significance level to create a Type I error rate of .05 does not sufficiently adjust the nominal significance level for the bootstrapping methods in order to result in their having a Type I error rate of .05. Doing another analysis to make such an adjustment for the bootstrapping methods would have required an extensive amount of additional computer time; consequently, a different approach was used. Three representative cases were selected from the second power analysis: the [20, 2] cases with variance ratios of 5:1 and 10:1 and the [50, 50] case with variance ratio of 1:5. The [20, 2] cases were selected because one sample size was large and the other small and $\sigma_1 > \sigma_2$. The [50, 50] case was selected because it was representative of the cases for which both sample sizes were large and equal. The third power analysis was then conducted as follows.

The effect sizes were calculated for the [20, 2] and [50, 50] cases for which homogeneity of variance had been maintained (the $\sigma_1 = \sigma_2$ cases). Effect size is defined as the follows:

$$\omega^2 = \frac{(\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

where ω^2 = the effect size, σ_i^2 = the variance for population i, and n_i = the sample size taken from population i. The difference in effect size between the various cases for given sample sizes is in the different values for the population variances. To maintain effect size in the cases where the homogeneity of variance assumption had been violated, the denominator for the above equation was recalculated by setting it equal to the appropriate homogeneity of variance value and solving for the variance values, while maintaining the variance ratio. The three cases were then run again with the adjusted variances.

Kulkarni's BNP program (1993) outputs Type I error rates and power based on four types of confidence interval. The output relevant to this study is discussed in detail, but the entire output can be found in Appendix B for Phase One (Type I error rates) and in Appendices C, D, and E for Phase Two (power), and a short description of the statistical tests conducted by BNP but not utilized in this study can be found in Appendix F.

Phase I - Type I Error Rates:

The software package DATASIM (Bradley, 1993) was used to generate 31 conditions (pairs of data sets) with 2000 replications of each condition (pair) for a total of 62,000 data sets of various sizes and degrees of heterogeneity of variance that were used as initial samples. The data sets were normally distributed with means of zero. The DATASIM package was selected because it is specifically designed for use in theoretical

statistical research and is easy to learn to use. The software has been tested extensively and has been found to perform well as a simulation tool (Bradley, Senko, & Stewart, 1990). Empirical distributions for F, t, z, r, and χ^2 obtained using DATASIM were found to be close to their respective theoretical distributions. And, DATASIM analyses of Type I error rates were found to replicate earlier empirical and theoretical studies. In cases where DATASIM's analysis did not agree with standard published tables, research has shown that the DATASIM values were more accurate (Bradley, Senko, & Stewart, 1990). DATASIM was also one of the packages used by Kulkarni (1993), and using it again made replication of that work straight forward. A more detailed discussion of DATASIM can be found in Kulkarni (1993) and Bradley (1993), and the exact DATASIM program implementation used in this study can be found in Appendix G. The characteristics of the data sets generated are shown in Table 1.

The [5, 3] and [100, 20] cases (indicating cases where $n_1 = 5$ and 100, respectively and $n_2 = 3$ and 20, respectively) were chosen as replication points from Kulkarni (1993). The former gave an indication of the performance of Student's t-test, Welch's t'-test, and both bootstrapping methods in the case of two small unequal samples; the latter gave an indication of the performance of the same statistical tests in the case of two large unequal samples. In addition, the [100, 20] case most nearly matches the [200, 20] cases which were used in this study as measures of the asymptotic performance of the [20, 2] and [50, 5] cases, as all three have a sample size ratio of 10:1. The 5:1 sample size ratio of the [100, 20] case was expected to produce Type I error rates that tend toward the same direction as the [50, 5] and [200, 20] cases. The [50, 5] cases were selected to explore

the performance of the combination of one large and one small sample with respect to actual Type I error rates for the above mentioned statistical tests. Finally, the [2, 2], [5,5], [50, 50], and [200,200] cases were selected to explore performance when samples are equal and small or equal and large, with respect to actual Type I error rates for the same statistical tests. The ratios of the degree of heterogeneity of variance were selected to be $\sigma_1^2:\sigma_2^2$ = 1:10, 1:5, 1:1, 5:1, and 10:1. These ratios were selected because they spanned the range from a large degree of heterogeneity to homogeneity of variance and because they were used in the Kulkarni work, facilitating comparison between the two studies.

Student's t-test and Welch's t'-test were conducted as indications of how parametric measures compare to bootstrapping with respect to Type I error rates. Two bootstrapping procedures were utilized, both of which were used by Kulkarni (1993) and were adapted from Lunneborg (1987). Lunneborg based the bootstrapping algorithms on the linear model upon which the ANOVA depends. For the purposes of this study, the general model can be expressed as $Y_{i,j} = \mu_i + \epsilon_{i,j}$ where $Y_{i,j}$ = the jth observation from sample i, μ_i = the mean of the population from which sample i is taken, and $\epsilon_{i,j}$ = the error present in the jth observation of sample i (Keppel, 1991; Lunneborg, 1987). As a population parameter, μ_i is never known, but the best estimate available for μ_i is $\hat{\mu}_i = \bar{Y}$, the sample mean. The value of an error score, $\epsilon_{i,j}$, can be estimated by $Y_{i,j} - \hat{\mu}_i$. The error score for each observation can be thought of has having been drawn from an error pool. In the case of the ANOVA, homogeneity of variance is assumed, and the population error pool is identical for all samples. This means that the error scores for the samples can be combined into one large pool as the best estimate of this common

population error pool. In cases of heterogeneity of variance where the error pools are not identical for all samples, each sample requires a separate error pool. The former case is referred to as a pooled error case and uses an error term related to the error term for a Student's t-test, whereas the latter case is referred to as an unpooled error case and uses an error term related to the error term for a Welch's t'-test.

Lunneburg (1987) made use of this model to devise two bootstrapping algorithms, one with a pooled error term and one with an unpooled error term, both of which were used in this study. Consider again the following example. Let $Y_1 = \{-4, -2, 1, 3, 7\}$ be a sample of size $n = 5$ from a population with $\mu_1 = 0$ and $\sigma_1^2 = 25$. Let $Y_2 = \{1, 2, 3, 4, 4, 6, 7\}$ be a sample of size $n = 7$ from another population with $\mu_2 = 0$ and $\sigma_2^2 = 25$. The sample mean for the first sample is $\overline{Y_1} = 1.00$, and the sample mean for the second sample is $\overline{Y_2} = 3.86$. The difference between the population means is $\delta = 0 - 0 = 0$, whereas the difference between the sample means is $d = 1.00 - 3.86 = -2.86$.

The pooled bootstrap procedure is then conducted as follows. The error pool is created by calculating the differences between each observation and its sample mean. Since homogeneity of variance is assumed, all error scores are placed in a single error pool. The error scores for this example are:

Y_1 errors:		Y_2 errors:	
	-4.00 - 1.00 = -5.00		1.00 - 3.86 = -2.86
	-2.00 - 1.00 = -3.00		2.00 - 3.86 = -1.86
	1.00 - 1.00 = 0.00		3.00 - 3.86 = -0.86
	3.00 - 1.00 = 2.00		4.00 - 3.86 = 0.14
	7.00 - 1.00 = 6.00		4.00 - 3.86 = 0.14

$$6.00 - 3.86 = 2.14$$

$$7.00 - 3.86 = 3.14$$

The error pool is the set of the resultant $(n_1 + n_2) = 12$ error scores, {-5.00, -3.00, -2.86, -1.86, -0.86, 0.00, 0.14, 0.14, 2.00, 2.14, 3.14, 6.00}. This error pool along with the means of the sample observations, $\frac{\sum_{j=1}^{n_1} Y_{1,j}}{n_1} = 1.00$ and $\frac{\sum_{j=1}^{n_2} Y_{2,j}}{n_2} = 3.86$ are used to create a large number (1,000 in this study) of bootstrap replications.

A set of $n_1 = 5$ and a set of $n_2 = 7$ error scores are drawn with replacement from the pool of twelve error scores and their means calculated. One such drawing might result in the following two error means: $[(-0.86) + 0.00 + 2.14 + 6.00 + (-3.00)]/5 = 0.86$ and

$$[2.00 + 2.00 + 3.14 + (-5.00) + (-2.86) + (-0.86) + 0.14]/7 = -0.21$$

These values are then used to estimate bootstrap means for the two populations as well as their difference. $\overline{Y^*_{1,j}} = 1.00 + 0.86 = 1.86$ and $\overline{Y^*_{2,j}} = 3.86 + (-0.21) = 3.65$. A bootstrapping replication, d_j^*, is defined as the difference between the two jth (where j = 1, 2, ..., 1000) bootstrapping means, so $d_j^* = 1.86 - 3.65 = -1.79$.

A 95% confidence interval about each set of 1,000 values for d_j^*, the difference between the bootstrapping means, and the actual difference between the sample means, d, is found in order to calculate the Type I error rate. The Type I error rate is found by counting the number of intervals that do not include zero within their limits and dividing by 1,000 replications. If, for example, 61 of the confidence intervals did not include zero, then the Type I error rate would be $\frac{61}{1000} = .061$.

Consider now the unpooled bootstrap procedure. The error pool is again created by calculating the differences between each observation and its sample mean. Since

homogeneity of variance is not assumed, the error scores from each sample are placed in a separate error pool. Once again, the error scores for this example are:

Y_1 errors:	Y_2 errors:
-4.00 - 1.00 = -5.00	1.00 - 3.86 = -2.86
-2.00 - 1.00 = -3.00	2.00 - 3.86 = -1.86
1.00 - 1.00 = 0.00	3.00 - 3.86 = -0.86
3.00 - 1.00 = 2.00	4.00 - 3.86 = 0.14
7.00 - 1.00 = 6.00	4.00 - 3.86 = 0.14
	6.00 - 3.86 = 2.14
	7.00 - 3.86 = 3.14

The error pools are the two sets of size $n_1 = 5$ and $n_2 = 7$ resultant error scores, {-5.00, -3.00, 0.00, 2.00, 6.00} and {-2.86, -1.86, -0.86, 0.14, 0.14, 2.14, 3.14}, respectively. These error pools along with the means of the sample observations, $\frac{\sum_{j=1}^{n_1} Y_{1,j}}{n_1} = 1.00$ and $\frac{\sum_{j=1}^{n_2} Y_{2,j}}{n_2} = 3.86$ are used to create a large number (1,000 in this study) of bootstrap replications.

A set of $n_1 = 5$ and a set of $n_2 = 7$ error scores are drawn with replacement from the corresponding error pools and their means calculated. One such drawing might result in the following two error means: [(-5.00) +0.00 + 0.00 + 6.00 + (-3.00)]/5 = -0.40 and [3.14 + 3.14 + 2.14 + (-2.86) + (-2.86) + (-1.86) + 0.14]/7 = 0.14

These values are then used to estimate bootstrap means for the two populations as well as their difference. $\overline{Y^*_{1,j}} = 1.00 + (-0.40) = 0.60$ and $\overline{Y^*_{2,j}} = 3.86 + 0.14 = 4.00$. Let d_j^* be the difference between the two jth (where j = 1, 2, ..., 1000) bootstrapping means, so $d_j^* = .60 - 4.00 = -3.40$.

A 95% confidence interval about each set of 1,000 values for d_j^*, the difference between the bootstrapping means and the actual difference between the sample means, d, is found in order to calculate the Type I error rate. The Type I error rate is found by counting the number of intervals that do not include zero within their limits and dividing by 1,000 replications. If, for example, 61 of the confidence intervals did not include zero, then the Type I error rate would be $\frac{61}{1000} = .061$.

Two thousand pairs of sample data sets were bootstrapped (each pair was bootstrapped 1,000 times) in order that the sampling distribution of the empirical proportion of Type I errors at the .05 level would approach normality. The sampling distribution of a proportion is binomially distributed with $p = \alpha_n = .05$ and $q = .95$. According to Glass & Hopkins (1996), a p value of .05 requires a sample size of at least 1400 in order for $(p - 1.96\alpha_p < p < p + 1.96\alpha_p)$, the 95% confidence interval for a standard normally distributed statistic, to accurately reflect the 95% confidence interval for the binomial case. Using the standard method above for finding symmetric confidence intervals, with $p = .05$ and $q = .95$, the 95% confidence interval is (.0404, .0596) with a width of .0191. Using the Ghosh method for bias correction in confidence intervals for a proportion of a population (Ghosh, 1979), the 95% confidence interval was calculated again. The resulting interval was (.0413, .0604) with a width of .0192. That these two methods produced close to the same result, suggests that if 2000 bootstrap simulations were used, the sampling distribution of Type I error rates should approach normality. This, in turn, means that 95% confidence intervals could be calculated using the percentile method without having to be concerned with making a correction for bias in the intervals.

Consequently, 2000 bootstrap simulations for each of the 31 cases were used because that should be large enough to suggest that the sampling distribution of the proportion would be approaching normality. Normality of the sampling distribution should, in turn, make the percentile confidence intervals similar to the bias-corrected intervals because a Normal distribution is symmetric around both its median and mean.

Actual Type I error rates were found by calculating the number of estimated differences between the means that were outside 95% confidence intervals for the difference between the means and dividing that value by 1,000. These bootstrapping confidence intervals were determined using the percentile method. This, in turn, yielded a distribution of 2,000 error rates.

The bootstrapping procedure was conducted using the software package, BNP, written by Kulkarni (1993). This package calculates Type I error rates for Student's t-test, the pooled and unpooled bootstrapping cases, and for certain values of n_1 and n_2 (i.e. $n_1 = 5$ and $n_2 = 3$ or $n_1 = 7$ and $n_2 = 7$) the Robust Rank Order test. Since the DATASIM simulation package is also capable of conducting Welch's t'-test in addition to Student's t, both tests were run using DATASIM. This provided a check to make sure that BNP was calculating Student's t properly, thereby giving more confidence in BNP's other calculations. A detailed discussion of BNP can be found in Kulkarni (1993), and a complete discussion of DATASIM's hypothesis testing capability can be found in Bradley (1988).

The actual computer implementation consisted of running DATASIM to generate the 31 conditions with 2000 replications for each condition for a total of 62,000 data sets

and calculating the actual Type I error rates for Student's t and Welch's t'. The resultant data sets were used as input into BNP (Kulkarni, 1993) which calculated the bootstrapping and Student's t Type I error rates. All simulations were done on a Gateway 2000 100Mh Pentium computer.

Phase II - Power:

The software package DATASIM (Bradley, 1993) was used to generate 31 conditions (pairs of data sets) with 2000 replications of each condition (pair) for a total of 62,000 data sets of various sizes and degrees of heterogeneity of variance to be used as initial samples, although for the power analysis the data sets were normally distributed with unequal means. The exact DATASIM program implementation used in this study can be found in Appendix G, and the characteristics of the data sets generated are shown in Table 2.

The data sets selected for power analyses were identical to those used in Phase One of the study with one exception. In order to investigate power, these data sets were generated with unequal means. One data set from each pair was generated with a mean of zero; the other was generated with a nonzero mean which produced a power of approximately .70 for Student's t-test when the population variances for the two sets were homogeneous.

Generally, power is conceptualized as the probability of finding existent treatment effects when the probability of making a Type I error would be .05 (or .01) had there not been treatment effects. When the homogeneity of variance assumption is violated and means are equal, α_a is not always equal to α_n, as suggested in Phase One of this study.

When investigating power a comparison needed to be made between cases where the ANOVA assumptions were violated and those where the assumptions were not violated. In order to make such a comparison, it was necessary to find the α_n which produced an $\alpha_a = .05$ for each case where the homogeneity of variance assumption was violated but the means were equal. Using this adjusted α_n in the power analysis made it possible to compare cases from Phase One with the corresponding cases in Phase Two. The value of the nonzero mean was determined by Monte Carlo simulations.

Further, since the α_n which was adjusted for the power analysis was for Student's t-test, the results were still ambiguous. Consequently, three cases (the [20, 2] cases with variance ratios of 5:1 and 10:1 and the [50, 50] case with variance ratio of 1:5) were run a third time. In this final analysis, rather than adjusting α_n, the values of σ_1^2 and σ_2^2 were adjusted so as to maintain a constant effect size. The values of σ_1^2 and σ_2^2 were derived mathematically and were based on the effect sizes of the [20, 2] case with variance ratio 1:1 (for the two [20, 2] cases) and of the [50, 50] case with variance ratio 1:1 (for the [50, 50] case).

The same statistical tests used in Phase One of the study were also used in Phase Two. Student's t, Welch's t', and two bootstrapping procedures, one using a pooled error term and the other using an unpooled error term, were utilized. Actual power was found by calculating the number of actual differences between the bootstrapping means, d_j^*, that were outside the 95% or $(1.00 - \alpha_n)\ x\ 100\%$ confidence interval for the difference between the means, depending on whether or not an adjusted α_n was being used. Bootstrapping confidence intervals were determined using the percentile method, as done

in Phase One. For specifics about Student's t, Welch's t', or the bootstrapping procedures, the reader is referred to the Phase One methodology of this study.

Results and Discussion

Phase I - Type I Error Rates:

The purpose of this phase of the study was to replicate and expand upon the work of Kulkarni (1993). Kulkarni examined the effects upon α_a of various degrees of heterogeneity of variance in conjunction with differing combinations of sample size when using two forms of the bootstrapping procedure, the parametric Student's t-test, and the nonparametric Robust Rank Order test. The two bootstrapping procedures included one test that used a pooled error term and one that used an unpooled error term (Kulkarni, 1993) in the calculation of the standard error of the statistic. In addition to the two bootstrapping procedures and Student's t-test, the present study examined the performance of Welch's t'-test, a parametric t-test corrected for heterogeneous variances (Bradley, 1993; Glass & Hopkins, 1996). The Robust Rank Order test was not of interest in this study and is discussed only in Appendices B and F. Kulkarni did not examine cases where sample sizes are unequal but homogeneity of variance is maintained, nor did she look at large equal sample sizes or cases where one sample is large ($n \geq 20$, in this study) and the other small ($n \leq 5$, in this study). This study addressed these omissions and also tried to validate further Kulkarni's (1993) results through replication. The results of Kulkarni's BNP program that are pertinent to this study can be found in Table 3, while the complete computer output can be found in Appendix B.

The actual Type I error rates found for the [5, 3] and [100, 20] replication points

were not significantly different from those of Kulkarni (1993), adding validity to Kulkarni's results. In the [5, 3] case where the samples were small and unequal and $\sigma_1^2 < \sigma_2^2$, none of the tests performed well; all of the actual Type I error rates were outside the acceptable interval of (0.040, 0.060) and too large. The pooled bootstrapping technique resulted in Type I error rates that were nearly twice as large as those obtained by Student's t-test (which also uses a pooled error term), and the unpooled bootstrapping technique resulted in Type I error rates that were more than three times greater than Welch's t'-test (which also uses an unpooled error term). While both parametric tests (t and t') resulted in inflated Type I error rates, Welch's t'-test produced a Type I error rate less than half the size of Student's t-test. The fact that Type I error rates were inflated is not entirely surprising because the larger sample was paired with the smaller variance (Glass & Hopkins, 1996; Hsu, 1938; Kulkarni, 1993). Error is underestimated with such pairings, yielding overly narrow confidence intervals which do not include zero as often as they should.

In the [100, 20] case, where the samples were large and unequal and $\sigma_1^2 < \sigma_2^2$, Welch's t'-test (using an unpooled error term) produced a Type I error rate of .046, the closest to the nominal of .05. The unpooled bootstrapping method produced a Type I error rate nearly one and a half times greater than that for Welch's t'. The pooled bootstrapping method and Student's t-test (both of which use a pooled error term) produced Type I error rates approximately four times the nominal rate of .05.

When sample sizes were unequal and large with $n_1 > n_2$ and $\sigma_1^2 < \sigma_2^2$ (the [100, 20] and [200,20] cases), Welch's t'-test which uses an unpooled error term, resulted in actual

Type I error rates closest to the nominal (.046 for the [100, 20] case and .045 for the [200, 20] case). For the [100, 20] case, Student's t-test and the pooled bootstrap, both of which use a pooled error term, resulted in Type I error rates (.226 and .233, respectively) approximately five times higher than that of Welch's t'-test (.046) which uses an unpooled error term. The Type I error rate for the pooled bootstrap (.233) was also more than three times greater than that of the unpooled bootstrap Type I error rate of .067, a Type I error rate nearly one and a half times higher than that of Welch's t' though only slightly inflated relative to α_n. For the [200, 20] case, Student's t-test resulted in Type I error rates from six to nine times greater than those of Welch's t'-test, which produced Type I error rates of .047 and .045 for the variance ratios of 1:5 and 1:10, respectively (.276 and .389, respectively). The pooled bootstrap resulted in Type I error rates four to six times greater than those of the unpooled bootstrap, which were .067 and .063 for the variance ratios of 1:5 and 1:10, respectively (.275 and .389, respectively). In both cases, the tests with unpooled error terms produced Type I error rates much closer to $\alpha_n = .05$ than did the tests with pooled error terms. In all four cases, the larger sample was paired with the smaller variance, so one would expect the Type I errors to be inflated in comparison to the same samples with the variances reversed (Hsu, 1938; Kulkarni, 1993).

When both samples were small and equal (the [2, 2] and [5, 5] cases), the parametric tests resulted in more accurate Type I error rates (i.e. $\alpha_a \approx \alpha_n$) regardless of the variance ratio than did the bootstrapping methods. In the [5, 5] cases, Welch's t' produced Type I error rates closest to $\alpha_n = .05$ in two of the three cases. In the remaining case (with a 1:5 variance ratio) where the Type I error rate for Student's t was closer than

Welch's t' to $\alpha_n = .05$, the difference between the two methods was only .011 and both were in the acceptable interval of (.04, .06). In the [2, 2] cases, Welch's t' produced Type I error rates closest to $\alpha_n = .05$ in only one case (with a variance ratio of 1:10). In the two cases where the Type I error rate for Student's t was closer to $\alpha_n = .05$ than was Welch's t', the difference between the two methods was .025 for the 1:5 variance ratio and .053 for the 1:10 variance ratio. When both samples were large and equal (the [50, 50] and [200, 200] cases), both parametric tests and both bootstrapping methods resulted in actual Type I error rates not significantly different from the nominal level of .05 regardless of the variance ratio (i.e. α_a within $\pm$ .01 of α_n).

When one sample was large and the other small with $n_1 > n_2$ and $\sigma_1^2 < \sigma_2^2$ (the first two [20, 2] and [50, 5] cases in Table 3), Welch's t' in the [50, 5] cases was the only test to result in Type I error rates within $\pm$ .01 of α_n. But, while the unpooled bootstrap method resulted in Type I error rates between three to four times those of Welch's t'-test, Student's t-test and the pooled bootstrap resulted in Type I error rates from two and a half to seven times greater.

When one sample was large and the other small with $n_1 > n_2$ and $\sigma_1^2 = \sigma_2^2$ (the third [20, 2] and [50, 5] cases listed in Table 3), the unpooled bootstrap method resulted in a Type I error rates significantly larger than $\alpha_n = .05$ (.316 and .141, respectively). Student's t-test produced Type I error rates not significantly different from $\alpha_n = .05$ in both cases, .049 for [20, 2] and for [50, 5]. The Type I error rates for Welch's t' were slightly larger, but still within $\pm$ .01of α_n, in the [50, 5] case (.057) but approximately twice the nominal value for the [20, 2] case (.118). The pooled bootstrap produced Type

I error rates of .074 for [20, 2] and .059 for [50, 5]. When $n_1 > n_2$ and $\sigma_1^2 > \sigma_2^2$ for the same sample sizes (the last two [20, 2] and [50, 5] cases listed in Table 3), both Student's t-test and the pooled bootstrap produced Type I error rates that were significantly smaller than the nominal alpha of .05. This was not surprising as the larger samples had the larger variances (Hsu, 1938; Kulkarni, 1993). Welch's t'-test performed well with a Type I error rate within ± .01 of nominal for the [50, 5] cases, with only slightly inflated Type I error rates in the [20, 2] cases (.074 and .065). The unpooled bootstrap was somewhat inflated in the [50, 5] cases (.101 and .083), but in the [20, 2] cases, the unpooled bootstrap produced Type I error rates were two and a half to three times greater than nominal (.163 and .124).

Phase II - Power:

The purpose of this phase of the study was to better understand the power of the bootstrapping procedure relative to that of Student's t-test and Welch's t'-test. All tests were run at the $\alpha_n = .05$ significance level, and the results of the BNP program pertinent to the initial analysis for this study can be found in Table 4. The complete BNP computer output can be found in Appendix C.

The initial analysis as seen in Table 4 appears to suggest that all four tests have equivalent power (approximately .70) when sample variances are equal. The exceptions to this are for Welch's t'-test in the [50, 5] case ($(1 - \beta) = .5525$), the [20, 2] case ($(1 - \beta) = .3715$), and the [2, 2] case ($(1 - \beta) = .4040$). This analysis appears to further suggest that when both sample sizes are small ($n \leq 5$, in this study) or when both sample sizes are large but equal (the [50, 50] and [200, 200] cases) and the variance ratios are

1:10 or 1:5, the bootstapping methods have power equal to or somewhat higher than the parametric Student's t and Welch's t'. In the remaining cases, where sizes were unequal and the variance ratios were 5:1 or 10:1, the analysis appears to suggest that Student's t and the pooled bootstrapping method result in lower power than Welch's t' and the unpooled bootstrapping method. But, looking back at Phase One of the study, the lack of standardization for α_a in the Type I analysis indicates that an unambiguous interpretation of this initial power analysis is impossible.

The emerging patterns between the Type I and the power analyses make it apparent that relative power among the cases and tests can not be determined. The problem is that when power is discussed, it is done with the understanding that had there been no differences between the population means, the Type I error rate would be .05. Many of these data sets did not result in a Type I error rate of .05 in Phase One of this study. Consequently, an adjustment was needed. The value for α_n in the Phase One data was altered using Monte Carlo simulations until each case resulted in a Type I error rate of approximately $\alpha_a = .05$ for Student's t-test. This new α_n was then used as the nominal alpha in a new power analysis. By using this adjustment, it was possible to better compare the relative power of the four different tests used in the study. The results of the BNP program pertinent to this study for the power analyses done with an adjusted alpha level can be found in Table 5. The complete BNP computer output can be found in Appendix D.

The nine cases with homogeneity of variance maintained were not rerun with an adjusted alpha because the Type I error rate analysis indicated that these cases had a Type

I error rate of .05. Other cases which also had a Type I error rate of .05 when the population means were equal were the [5, 5] case with a variance ratio of 1:5, the [50, 50] cases with variance ratios of 1:10 and 1:5, and the [200, 200] cases with variance ratios of 1:10 and 1:5. The initial power results for these cases are reprinted in Table 5. Further, an additional seven cases were not run with the adjusted alpha because the required alpha would be so small ($\alpha_n \leq .001$) that 100% confidence intervals would have to be used in calculating power. These cases are denoted in Table 5 by the statement "not run."

Of the four cases where both sample sizes were small, all power results either stayed the same or went down as a result of the adjustment in α_n. As with the unadjusted power analysis, Student's t and Welch's t' appeared to have lower power than the two bootstrapping procedures.

Of the two cases where one sample was large and the other small (the [20, 2] and [50, 5] cases), only the cases with variance ratios of 5:1 and 10:1 were re-analyzed. The same pattern emerged as did in the initial power analysis. Student's t and the pooled bootstrap method appeared to have comparable power, and Welch's t' and the unpooled bootstrap method appeared to have comparable power. The former two tests appeared to result in lower power than the latter two tests, and in both situations the bootstrapping methods appeared to have the higher power. Further, all power values increased from the initial unadjusted power analysis.

Of the two cases where both sample sizes were large but unequal (the [100, 20] and [200, 20] cases), only two of the [200, 20] cases were re-analyzed. Those were the cases with variance ratios of 5:1 and 10:1. Here again, the pattern found in the original

power analysis was repeated; the power of Student's t appeared to be comparable to that of the pooled bootstrapping method, and the power of Welch's t' appeared to be comparable to that of the unpooled bootstrapping method. In all cases, power nearly doubled from the initial power analysis, and the unpooled bootstrapping method appeared to result in power equivalent to or higher than that found when the variances were homogeneous.

Again, the emerging patterns between the two power analyses make it apparent that relative power among the cases and tests can not be fully determined. The problem is that α_n was only adjusted for Student's t-test. It should also have been adjusted for the bootstrap methods. Since that would require a great deal of addition computer time, a different approach was attempted. In this final analysis, three representative cases were run a third time. This time the effect size was maintained. The results of the BNP program pertinent to this study for the power analyses done with the effect size maintained can be found in Table 6. The complete BNP computer output can be found in Appendix E.

In both [20, 2] cases, the power of Student's t and the pooled bootstrapping method increased, relative to the power of the same tests in the second (adjusted α_n) power analysis. The [20, 2] case with variance ratio of 10:1 resulted in a power value of .7390 and .8125 for Student's t-test and the pooled bootstrapping method, respectively, compared to power values of .0570 and .0910 for the same tests in the previous analysis. The [20, 2] case with variance ratio of 5:1 resulted in power values of .2370 and .3260 for Student's t-test and the pooled bootstrapping method, compared with power values of

.0475 and .0765 for the previous analysis. In neither of the [20, 2] cases was it possible to make a comparison with the unpooled bootstrapping method because it had not been possible to obtain Type I error rates of approximately $\alpha_a = .05$ in the earlier analyses. In the [50, 50] case, the power of all four tests increased relative to the power of the same tests in the second (adjusted α_n) power analysis. This analysis resulted in power values of .6920, .6885, .7085 and .7065 for Student's t, Welch's t', and the pooled and unpooled bootstrapping methods, respectively, compared with power values of .1750, .1720, .1895 and .1870 for the same tests in the previous analysis.

General Discussion:

The overall purpose of the present study was to understand the power of the bootstrap statistic relative to Student's t test. Two forms of the bootstrapping statistic were used in order to assess its performance. One form of the bootstrapping method used a pooled error term in the calculation of the standard error of the mean, while the other form used an unpooled error term in the calculation of the standard error of the mean. Student's t also uses a pooled error term in the calculation of the standard error of the mean, making it a logical choice for comparison to the pooled bootstrapping method. Welch's t' is similar to Student's t but uses an unpooled error term in the calculation of the standard error of the mean, making it a logical choice for comparison to the unpooled bootstrapping method.

In order to understand power, a Type I error rate analysis was conducted first to establish how each of the four tests performed when there was no difference between the means. The Type I error rate analysis was followed by three different power analyses. By

definition, power is the probability that a test correctly rejects the null hypothesis when there is a difference in the means, while assuming that had there not been a difference the Type I error rate would be $\alpha_a = .05$. Consequently, the first power analysis was used to establish the power of Student's t, Welch's t', and both bootstrapping methods for those cases which had resulted in a Type I error rate of approximately $\alpha_a = .05$ (i.e. $.04 \leq \alpha_a \leq .06$). The second power analysis was used to establish the power of those cases where the nominal alpha could be adjusted such that $\alpha_a = .05$ in the Type I error analysis. Nominal alpha was adjusted only for Student's t test, so the extent to which the second power analysis could be generalized was limited. Finally, the third power analysis considered the possibility of using effect size, rather than Type I error rates, as a baseline for the comparison of the relative power of the different tests.

When all of the assumptions for the analysis of variance are met, Student's t test was found to result in the highest power, which is in agreement with the researcher's hypothesis. As a parametric test, Student's t should be more powerful than the nonparametric bootstrapping methods when all ANOVA assumptions are met. Welch's t' test generally performed well also in cases where homogeneity of variance was maintained. The exception to this was for the [2, 2] and [20, 2] cases, for which the Type I error rates were deflated and inflated, respectively. The pooled bootstrap method resulted in power equivalent to the parametric cases when both sample sizes were large (and in the [50, 5] case) and homogeneity of variance was maintained, whereas the unpooled bootstrapping method only resulted in power equivalent to the parametric cases in the [50, 50] and [200, 200] cases, again when homogeneity of variance was maintained.

Given the pattern of Type I error rates and power results, the researcher suspects that the unpooled bootstrapping method also performs well in the [100, 20] and [200, 20] cases. Thus, when all ANOVA assumptions are met, Student's t appears to be the best choice of statistical tests. It is the simplest test to use and is powerful.

When homogeneity of variance is not maintained, power appears to go down for all four tests. When using Type I error rates as a baseline for establishing power and looking at cases where one sample size is large and the other small and with the large variance paired with the large sample, the pooled bootstrapping method results in slightly higher power than Student's t. Welch's t' resulted in power higher than Student's t in all cases except for the [50, 5] case with a large variance ratio (10:1). The reason for this exception is not clear at this point. In this same case, the pooled bootstrap performed slightly better than both parametric tests. When the variance ratio was moderately large (5:1), again for the [50, 5] case, the pooled bootstrap once again appears to result in slightly higher power than Student's t but significantly lower power than Welch's t'. While no test resulted in power near the desired .70, the bootstrap appears to be the best choice of tests. In none of these cases was it possible to make a comparison with the unpooled bootstrapping method because at no time did its Type I error rate reach $\alpha_a = .05$.

When both sample sizes were small, neither of the bootstrapping methods could be used for a comparison to Student's t because their Type I error rates were consistently too large. In fact, only in the [5, 5] cases and in one of the [5, 3] cases was a comparison between Student's t and Welch's t' possible. In those cases where a comparison was possible, Student's t resulted in slightly higher power, even in the case of heterogeneity of

variance; however, in no case with heterogeneity of variance did power results get beyond .2595, much lower than the desired .70.

When both sample sizes were large but unequal and the homogeneity of variance assumption was violated, it was possible to make more comparisons, although some of the researcher's conclusions are based on the patterns of Type I error rates and power results. In the cases where the smaller sample was paired with the larger variance, it was not possible to compare Student's t to the pooled bootstrapping method because the Type I error rates for both tests were too large. In fact, the Type I error rates were so large that in order to make an adjustment so that $\alpha_a = .05$, nominal alpha would have to be set at zero. Welch's t', on the other hand, had power substantially lower than desired at between .1200 and .1300. In the cases where the larger sample was paired with the larger variance (occurring only for the [200, 20] sample sizes), Welch's t' resulted in slightly lower power than Student's t. The pooled bootstrapping method resulted in slightly higher power than Student's t, and based on the patterns of Type I error rates and power results the researcher suspects that the unpooled bootstrapping method has power slightly higher than Welch's t' and slightly lower than Student's t and the pooled bootstrapping method. Consequently, the pooled bootstrap may be a good choice of tests under these circumstances.

When both sample sizes are large and equal and the homogeneity of variance assumption is violated, the pooled bootstrapping method resulted in power equal to or slightly higher than Student's t. Under the same conditions, the unpooled bootstrapping method resulted in power equal to or slightly higher than Welch's t'. In none of the cases

did any of the four tests result in power near the desired .70. All of the tests resulted in power between .1720 and .1965.

The attempt to use effect size, rather than Type I error rates, as a baseline for establishing power, suggests a potentially valuable tool in future work. The effect sizes for the three chosen cases with heterogeneity of variance were adjusted to be equal to the effect sizes of each case when homogeneity of variance was not violated. The fact that only three cases were considered using this technique, means that the results, while encouraging, are not conclusive.

The [50, 50] case with a moderate variance ratio (1:5) was selected because the standard power analysis resulted in all four tests having low power (between .1720 and .1965) even though their Type I error rates were approximately $\alpha_a = .05$ in the original Type I error rate analysis. The effect size analysis resulted in all four tests having power of approximately .70. Both bootstrapping methods resulted in slightly higher power than that of the parametric tests. This suggests that in situations where both samples are large and equal and homogeneity of variance is violated but the effect size matches that for when all ANOVA assumptions are met, the bootstrapping method is just as, if not slightly more, powerful as the parametric Student's t and Welch's t'.

The other two cases selected for the effect size analysis were the [20, 2] cases with moderate and large variance ratios and the large sample paired with the large variance. In both of these cases, power increased substantially from the original power analysis for each of the parametric test, with Welch's t' being the more powerful test. While it is not possible to make conclusive remarks about the bootstrapping methods, based on the

patterns of Type I error rates and power results the researcher suspects that the pooled and unpooled bootstrapping methods have power equivalent or slightly higher than Student's t and Welch's t', respectively.

There is still much work to be done before the bootstrap statistic is fully understood. In any future work, researchers should, in addition to adjusting nominal alpha rates, adjust variance levels as well in order to maintain effect sizes when trying to establish power. Thus far, it appears that the bootstrap works well when sample sizes are large. It is too soon to make conclusions about how the bootstrap works with small sample sizes because Type I error rates are extremely inflated, but the researcher anticipates that bootstrapping will not be as powerful with small samples.

Table 1

Various Sample Sizes and Ratios of Population Variances for Values chosen for Type I Error Rate Analysis

n_1	n_2	$\sigma_1^2:\sigma_2^2$	σ_1	σ_2
2	2	1:10	1.000000	3.162278
2	2	1:5	1.414214	3.162278
2	2	1:1	1.000000	1.000000
5	3	1:10	1.000000	3.162278
5	3	1:1	1.000000	1.000000
5	5	1:10	1.000000	3.162278
5	5	1:5	1.414214	3.162278
5	5	1:1	1.000000	1.000000
20	2	1:10	1.000000	3.162278
20	2	1:5	1.414214	3.162278
20	2	1:1	1.000000	1.000000
20	2	5:1	3.162278	1.414214
20	2	10:1	3.162278	1.000000

Table 1 continued

Various Sample Sizes and Ratios of Population Variances for Values chosen for Type I Error Rate Analysis

n_1	n_2	$\sigma_1^2:\sigma_2^2$	σ_1	σ_2
50	5	1:10	1.000000	3.162278
50	5	1:5	1.414214	3.162278
50	5	1:1	1.000000	1.000000
50	5	5:1	3.162278	1.414214
50	5	10:1	3.162278	1.000000
50	50	1:10	1.000000	3.162278
50	50	1:5	1.414214	3.162278
50	50	1:1	1.000000	1.000000
100	20	1:5	1.414214	3.162278
100	20	1:1	1.000000	1.000000
200	20	1:10	1.000000	3.162278
200	20	1:5	1.414214	3.162278
200	20	1:1	1.000000	1.000000

Table 1 continued

Various Sample Sizes and Ratios of Population Variances for Values chosen for Type I Error Rate Analysis

n_1	n_2	$\sigma_1^2:\sigma_2^2$	σ_1	σ_2
200	20	5:1	3.162278	1.414214
200	20	10:1	3.162278	1.000000
200	200	1:10	1.000000	3.162278
200	200	1:5	1.414214	3.162278
200	200	1:1	1.000000	1.000000

Table 2

Various Sample Sizes, Ratios of Population Variances for Values, and Means chosen for Power Analysis

n_1	n_2	$\sigma_1^2:\sigma_2^2$	σ_1	σ_2	μ_1	μ_2
2	2	1:10	1.000000	3.162278	0	
2	2	1:5	1.414214	3.162278	0	
2	2	1:1	1.000000	1.000000	0	4.80
5	3	1:10	1.000000	3.162278	0	
5	3	1:1	1.000000	1.000000	0	2.20
5	5	1:10	1.000000	3.162278	0	
5	5	1:5	1.414214	3.162278	0	
5	5	1:1	1.000000	1.000000	0	1.80
20	2	1:10	1.000000	3.162278	0	
20	2	1:5	1.414214	3.162278	0	
20	2	1:1	1.000000	1.000000	0	1.92
20	2	5:1	3.162278	1.414214	0	
20	2	10:1	3.162278	1.000000	0	

Table 2 continued

Various Sample Sizes, Ratios of Population Variances for Values, and Means chosen for Power Analysis

n_1	n_2	$\sigma_1^2:\sigma_2^2$	σ_1	σ_2	μ_1	μ_2
50	5	1:10	1.000000	3.162278	0	
50	5	1:5	1.414214	3.162278	0	
50	5	1:1	1.000000	1.000000	0	1.20
50	5	5:1	3.162278	1.414214	0	
50	5	10:1	3.162278	1.000000	0	
50	50	1:10	1.000000	3.162278	0	
50	50	1:5	1.414214	3.162278	0	
50	50	1:1	1.000000	1.000000	0	0.50
100	20	1:5	1.414214	3.162278	0	
100	20	1:1	1.000000	1.000000	0	0.62
200	20	1:10	1.000000	3.162278	0	
200	20	1:5	1.414214	3.162278	0	

Table 2 continued

Various Sample Sizes, Ratios of Population Variances for Values, and Means chosen for Power Analysis

n_1	n_2	$\sigma_1^2:\sigma_2^2$	σ_1	σ_2	μ_1	μ_2
200	20	1:1	1.000000	1.000000	0	0.59
200	20	5:1	3.162278	1.414214	0	
200	20	10:1	3.162278	1.000000	0	
200	200	1:10	1.000000	3.162278	0	
200	200	1:5	1.414214	3.162278	0	
200	200	1:1	1.000000	1.000000	0	0.25

Table 3

Actual Type I Error Rates when Ho is True and Nominal Significance is $\alpha = .05$ for Student's t-test, Welch's t'-test, and the Pooled and Unpooled Bootstrap tests

n_1	n_2	$\sigma_1^2:\sigma_2^2$	t	t'	PBP	UBP
2	2	1:10	.090	.037	.367	.409
2	2	1:5	.064	.035	.324	.364
2	2	1:1	.048	.026	.293	.326
5	3	1:10	.157	.071	.285	.243
5	3	1: 1	.055	.051	.143	.166
5	5	1:10	.069	.060	.131	.141
5	5	1:5	.055	.044	.131	.129
5	5	1: 1	.057	.051	.125	.129
20	2	1:10	.433	.106	.470	.432
20	2	1:5	.304	.125	.356	.411
20	2	1:1	.049	.118	.074	.316
20	2	5:1	.001	.074	.003	.163
20	2	10:1	.000	.065	.001	.124

Table 3 continued

Actual Type I Error Rates when Ho is True and Nominal Significance is $\alpha = .05$ for Student's t-test, Welch's t'-test, and the Pooled and Unpooled Bootstrap tests

n_1	n_2	$\sigma_1^2:\sigma_2^2$	t	t'	PBP	UBP
50	5	1:10	.395	.057	.394	.159
50	5	1:5	.301	.057	.319	.155
50	5	1:1	.049	.057	.059	.141
50	5	5:1	.000	.059	.000	.101
50	5	10:1	.000	.055	.000	.083
50	50	1:10	.051	.049	.055	.055
50	50	1:5	.045	.044	.051	.055
50	50	1:1	.052	.052	.057	.053
100	20	1:5	.226	.046	.223	.067
100	20	1:1	.048	.052	.053	.063
200	20	1:10	.389	.045	.389	.063
200	20	1:5	.276	.047	.275	.067
200	20	1:1	.052	.058	.055	.074

Table 3 continued

Actual Type I Error Rates when Ho is True and Nominal Significance is $\alpha = .05$ for Student's t-test, Welch's t'-test, and the Pooled and Unpooled Bootstrap tests

n_1	n_2	$\sigma_1^2:\sigma_2^2$	t	t'	PBP	UBP
200	20	5:1	.000	.054	.000	.065
200	20	10:1	.000	.046	.000	.054
200	200	1:10	.051	.050	.050	.051
200	200	1:5	.052	.052	.053	.053
200	200	1:1	.050	.050	.051	.050

Table 4

Actual Power when Nominal Significance is $\alpha = .05$ for Student's t-test, Welch's t'-test, and the Pooled and Unpooled Bootstrap tests (unadjusted α)

n_1	n_2	$\sigma_1^2:\sigma_2^2$	μ_1	μ_2	t	t'	PBP	UBP
2	2	1:10	0	4.80	.3030	.1515	.7645	.8140
2	2	1:5	0	4.80	.2600	.1290	.7450	.7835
2	2	1:1	0	4.80	.6780	.4040	.9980	.9990
5	3	1:10	0	2.20	.3155	.1375	.4850	.4210
5	3	1:1	0	2.20	.7120	.6085	.8935	.8990
5	5	1:10	0	1.80	.2120	.1710	.3520	.3625
5	5	1:5	0	1.80	.1865	.1655	.3220	.3270
5	5	1:1	0	1.80	.7040	.6700	.8440	.8420
20	2	1:10	0	1.92	.5875	.1325	.6245	.5605
20	2	1:5	0	1.92	.4590	.1565	.5135	.5225
20	2	1:1	0	1.92	.6985	.3715	.7695	.8970
20	2	5:1	0	1.92	.0265	.3095	.0540	.5525
20	2	10:1	0	1.92	.0140	.3850	.0360	.6300

Table 4 continued

Actual Power when Nominal Significance is $\alpha = .05$ for Student's t-test, Welch's t'-test, and the Pooled and Unpooled Bootstrap tests (unadjusted α)

n_1	n_2	$\sigma_1^2:\sigma_2^2$	μ_1	μ_2	t	t'	PBP	UBP
50	5	1:10	0	1.20	.5420	.1080	.5475	.2665
50	5	1:5	0	1.20	.4425	.1075	.4605	.2710
50	5	1:1	0	1.20	.7190	.5525	.7500	.7855
50	5	5:1	0	1.20	.0230	.2900	.0275	.4160
50	5	10:1	0	1.20	.0080	.4250	.0125	.5210
50	50	1:10	0	0.50	.1795	.1745	.1870	.1860
50	50	1:5	0	0.50	.1750	.1720	.1895	.1870
50	50	1:1	0	0.50	.7095	.7090	.7205	.7195
100	20	1:5	0	0.62	.3805	.1290	.3895	.1570
100	20	1:1	0	0.62	.7230	.7000	.7355	.7305
200	20	1:10	0	0.59	.5370	.1235	.5405	.1645
200	20	1:5	0	0.59	.4460	.1295	.4490	.1595
200	20	1:1	0	0.59	.7085	.6630	.7155	.7195

Table 4 continued

Actual Power when Nominal Significance is $\alpha = .05$ for Student's t-test, Welch's t'-test, and the Pooled and Unpooled Bootstrap tests (unadjusted α)

n_1	n_2	$\sigma_1^2:\sigma_2^2$	μ_1	μ_2	t	t'	PBP	UBP
200	20	5:1	0	0.59	.0215	.3015	.0205	.3330
200	20	10:1	0	0.59	.0065	.4570	.0085	.4755
200	200	1:10	0	0.25	.1965	.1965	.1940	.1965
200	200	1:5	0	0.25	.1745	.1740	.1805	.1785
200	200	1:1	0	0.25	.7055	.7055	.6960	.7065

Table 5

Actual Power for Student's t-test, Welch's t'-test, and the Pooled and Unpooled Bootstrap tests when Actual Significance is $\alpha = .05$ for Student's t-test (adjusted α)

n_1	n_2	$\sigma_1^2:\sigma_2^2$	μ_1	μ_2	α_n	t	t'	PBP	UBP
2	2	1:10	0	4.80	.03	.2595	.1215	.7485	.8140
2	2	1:5	0	4.80	.04	.2060	.0915	.7245	.7805
2	2	1:1	0	4.80	.05	.6780	.4040	.9980	.9990
5	3	1:10	0	2.20	.01	.1555	.0570	.3615	.3505
5	3	1:1	0	2.20	.05	.7120	.6085	.8935	.8990
5	5	1:10	0	1.80	.04	.1845	.1470	.3215	.3365
5	5	1:5	0	1.80	.05	.1865	.1655	.3220	.3270
5	5	1:1	0	1.80	.05	.7040	.6700	.8440	.8420
20	2	1:10	0	1.92	not run				
20	2	1:5	0	1.92	not run				
20	2	1:1	0	1.92	.05	.6985	.3715	.7695	.8970
20	2	5:1	0	1.92	.08	.0475	.3625	.0765	.5825

Table 5 continued

Actual Power for Student's t-test, Welch's t'-test, and the Pooled and Unpooled Bootstrap tests when Actual Significance is $\alpha = .05$ for Student's t-test (adjusted α)

n_1	n_2	$\sigma_1^2:\sigma_2^2$	μ_1	μ_2	α_n	t	t'	PBP	UBP
20	2	10:1	0	1.92	.11	.0570	.5210	.0910	.7140
50	5	1:10	0	1.20	not run				
50	5	1:5	0	1.20	not run				
50	5	1:1	0	1.20	.05	.7190	.5525	.7500	.7855
50	5	5:1	0	1.20	.09	.0565	.3975	.0685	.4995
50	5	10:1	0	1.20	.40	.4955	.8655	.5155	.8810
50	50	1:10	0	0.50	.05	.1795	.1745	.1870	.1860
50	50	1:5	0	0.50	.05	.1750	.1720	.1895	.1870
50	50	1:1	0	0.50	.05	.7095	.7090	.7205	.7195
100	20	1:5	0	0.62	not run				
100	20	1:1	0	0.62	.05	.7230	.7000	.7355	.7305
200	20	1:10	0	0.59	not run				

Table 5 continued

Actual Power for Student's t-test, Welch's t'-test, and the Pooled and Unpooled Bootstrap tests when Actual Significance is $\alpha = .05$ for Student's t-test (adjusted α)

n_1	n_2	$\sigma_1^2:\sigma_2^2$	μ_1	μ_2	α_n	t	t'	PBP	UBP
200	20	1:5	0	0.59	not run				
200	20	1:1	0	0.59	.05	.7085	.6630	.7155	.7195
200	20	5:1	0	0.59	.30	.3440	.6785	.3430	.6870
200	20	10:1	0	0.59	.39	.4670	.8420	.4765	.8415
200	200	1:10	0	0.25	.05	.1965	.1965	.1940	.1965
200	200	1:5	0	0.25	.05	.1745	.1740	.1805	.1785
200	200	1:1	0	0.25	.05	.7055	.7055	.6960	.7065

Table 6

Actual Power for Various Sample sizes when Effect Size is Maintained

n_1	n_2	σ_1	σ_2	μ_1	μ_2	α_n	t	t'	PBP	UBP
20	2	1.914854	0.733333	0	1.92	.08	.2370	.5475	.2360	.8670
20	2	1.348400	0.426401	0	1.92	.11	.7390	.9075	.8125	.9995
50	50	0.577350	1.290995	0	0.50	.05	.6920	.6885	.7085	.7065

References

Bradley, D. R. (1993). DATASIM. Desktop Press, Maine.

Bradley, D. R., Senko, W., and Stewart, F. A. (1990). Statistical simulation on microcomputers. Behavior Research Methods, Instruments, & Computers. 22, 236-246.

Buckland, S. T., (1984). Monte Carlo confidence intervals. Biometrics, 40, 811-817.

Diaconis, P. & Efron, B. (1983). Computer-Intensive methods in statistics. Scientific American, 248, 115-139.

Efron, B. (1979). Computers and the theory of statistics: Thinking the unthinkable. SIAM Review, 21(4), 460-480.

Efron, B. (1988). Bootstrap confidence intervals: Good or bad? Psychological Bulletin, 104(2), 293-296.

Efron, B. & Gong, G. (1983). A leisurely look at the bootstrap, the jackknife, and crossvalidation. The American Statistician, 37(1), 36-48.

Efron, B. & Tibshirani, R. J. (1991) . Statistical data analysis in the computer age. Science, 253, 390-395.

Efron, B. & Tibshirani, R. J. (1993). An introduction to the bootstrap. Great Britain: Chapman and Hall.

Ghosh, B.K. (1979). A comparison of some approximate confidence intervals for the binomial parameter. Journal of the American Statistical Association, 74(368), 894-900.

Glass, G. V. & Hopkins, K. D. (1996). Statistical methods in education and psychology (3rd. ed.). Needham Heights, MA: Allyn and Bacon.

Hsu, P. L. (1938). Contribution to the theory of Student's t-test as applied to the problem of two samples. Statistical Research Memoirs,2, 1-24.

Kulkarni, S. (1993). A comparison of type I error rates for the bootstrap contrast with the t test and the robust rank order test for various sample sizes and variances. Unpublished master's thesis, Lehigh University, Bethlehem, Pennsylvania.

Lunneborg, C. E. (1987). Bootstrap applications for the behavioral sciences (vol. 1). University of Washington.

Lunneborg, C. E. & Tousignant, J. P. (1985). Efron's bootstrap with application to the repeated measures design. Multivariate Behavioral Research, 20, 161-178.

Mendenhall, W. (1987). Introduction to probability and statistics (7th. ed.). Boston, Massachusetts: Duxbury.

Rassmussen, J. L. (1987). Estimating correlation coefficients: Bootstrap and parametric approaches. Psychological Bulletin. 101(1), 136-136.

Siegel, S. & Castellan, N. J., Jr. (1988). Nonparametric statistics (2nd. ed.). McGraw-Hill.

Strube, M. J. (1988). Bootstrap type I error rates for the correlation coefficient: An examination of procedures. Psychological Bulletin, 104(2), 290-292.

Appendix A

A simple type of confidence interval is based on the assumption that the bootstrapping sampling distribution of the statistic is unbiased and normally distributed. Such an "unadjusted" confidence interval, also called the symmetric method, is based on deviations from an (Strube, 1988). With sufficient bootstrap replications, about one thousand (Diaconis & Efron, 1983; Efron, 1979; Efron, 1988; Strube, 1988), the assumption of normality is warranted (Glass and Hopkins, 1996). In such situations, a confidence interval is defined as $(\hat{\theta} \pm z_{\alpha}\hat{s}_{\hat{\theta}})$ where $\hat{\theta}$ = the sample estimate of the statistic, z_{α} = the $100(\alpha)$ percentile of the normal distribution, and $\hat{s}_{\hat{\theta}}$ = the estimated standard error of $\hat{\theta}$.

The bias-corrected percentile method does not assume an unbiased median but does assume that the data are distributed such that there exists a monotone transformation of the data that is normal (Efron, 1988; Efron & Gong, 1983; Kulkarni, 1993; Strube, 1988). This method widens both confidence interval limits, but by different amounts based on the location of the median (Efron, 1988). To create a bias corrected confidence interval for a bootstrap estimate, one must first determine the percentile value of the sample estimate of the statistic, $\hat{\theta}$, within the bootstrap distribution, as well as the values of the standard scores $z_{\alpha/2}$ and $z_{(1-\alpha/2)}$ the normal distribution. The standard score below which the percentile value of $\hat{\theta}$ would lie in a normal distribution is defined as z_0. The upper and lower confidence interval limits are then calculated as $(2z_0 + z_{\alpha/2})$ and $(2z_0 + z_{(100-\alpha/2)})$, respectively.

Lastly, the minimum-width method is a special case of the percentile method (Efron & Gong, 1983; Kulkarni, 1993; Strube, 1988). To find a minimum-width confidence interval for a bootstrap statistic, the bootstrap replications, $d_j^* = (\overline{Y_1^*} - \overline{Y_2^*})$ where j = 1, 2, ..., B and $\overline{Y_i^*}$ is the mean of bootstrap sample i, are listed in ascending order. The positional difference is defined as $PD = z(1 - \alpha)B$. The differences for the bootstrap estimates are then found for estimates i and $(PD - 1 + i)$ where i = 1, 2, . , $(B - PD)$. The bootstrap estimates that result in the minimum width for the confidence interval are then defined as the interval's upper and lower limits. For example, suppose bootstrapping yields the following B = 15 bootstrap replications listed in ascending order.

{-1.40, -1.04, -0.56, -0.07, 0.06, 0.09, 0.16, 0.39, 0.54, 0.77, 0.99, 1.01, 1.04, 1.22, 1.50}

Suppose further, that the nominal alpha is .10. The positional difference is $PD = (1 - 0.10)(15) = 13.50$. The differences for bootstrap replications i and (PD - 1 + i) are then found. That is, one calculates the differences between bootstrap estimates 1 and 14 and between estimates 2 and 15.

For i = 1, | -1.40 - 1.22| = 2.62.

For i = 2, | -1.04 - 1.50| = 2.54.

Since 2.54 is smaller than 2.62, the minimum-width 90% confidence interval is (-1.04, 1.50).

Appendix B

The complete output for the Phase I simulation results are shown in Tables 7 through 15. The table column headings mimic those in Kulkarni (1993) and are as follows:

PBS - Pooled error Bootstrap Symmetric confidence interval method

PBP - Pooled error Bootstrap Percentile confidence interval method

PBB - Pooled error Bootstrap Bias-corrected confidence interval method

PBM - Pooled error Bootstrap Minimum-width confidence interval method

UBS - Unpooled error Bootstrap Symmetric confidence interval method

UBP - Unpooled error Bootstrap Percentile confidence interval method

UBB - Unpooled error Bootstrap Bias-corrected confidence interval method

UBM - Unpooled error Bootstrap Minimum-width confidence interval method

t - Empirically derived Student's t-test Type I error rates

TH - Theoretical t derived mathematically, Seheffe (1959) and Hsu (1938)

t' - Empirically derived Welch's t'-test Type I error rates

RR - Robust Rank Order test

All Type I error rates that fall between .041 and .060 are not significantly different at $\alpha_n = .05$. A discussion of the symmetric, bias corrected, and minimum width confidence intervals can be found in Appendix A, and a discussion of the Robust Rank Order test can be found in Appendix F.

Table 7

Actual Probability of Type I error for Sample sizes [2, 2] when H_0 is true and Nominal Significance is $\alpha = .05$

$\sigma_1^2:\sigma_2^2$	PBS	PBP	PBB	PBM	UBS	UBP	UBB	UBM	TH	t	t'	RR
1:10	.366	.367	.357	.387	.367	.409	.405	.238	.090	.037	.409	
1:5	.325	.324	.312	.346	.327	.364	.361	.954		.064	.035	.364
1:1	.294	.293	.275	.308	.296	.326	.329	.169		.048	.026	.326

Table 8

Actual Probability of Type I Error for Sample sizes [5, 3] when H_0 is true and Nominal Significance is $\alpha = .05$

$\sigma_1^2:\sigma_2^2$	PBS	PBP	PBB	PBM	UBS	UBP	UBB	UBM	TH	t	t'	RR
1:10	.276	.274	.276	.279	.231	.245	.240	.254	.145	.143	.073	.163
1:1	.141	.143	.146	.148	.158	.166	.167	.170		.055	.051	.079

Table 9

Actual Probability of Type I Error for Sample sizes [5, 5] when H_0 is true and Nominal Significance is $\alpha = .05$

$\sigma_1^2:\sigma_2^2$	PBS	PBP	PBB	PBM	UBS	UBP	UBB	UBM	TH	t	t'	RR
1:10	.133	.131	.133	.155	.133	.141	.140	.141		.069	.060	.099
1:5	.131	.131	.129	.138	.127	.129	.129	.133		.055	.044	.102
1:2	.125	.125	.129	.127	.123	.129	.130	.134		.057	.051	.101

Table 10

Actual Probability of Type I Error for Sample sizes [20, 2] when H_0 is true and Nominal Significance is $\alpha = .05$

$\sigma_1^2:\sigma_2^2$	PBS	PBP	PBB	PBM	UBS	UBP	UBB	UBM	TH	t	t'	RR
1:10	.476	.470	.469	.475	.381	.432	.434	.436		.433	.106	.423
1:5	.356	.356	.349	.367	.367	.411	.407	.412		.304	.125	.389
1:1	.073	.074	.079	.081	.300	.316	.319	.320		.049	.118	.252
5:1	.002	.003	.004	.002	.162	.163	.159	.166		.001	.074	.095
10:1	.001	.001	.001	.001	.124	.124	.128	.129		.000	.065	.072

Table 11

Actual Probability of Type I Error for Sample sizes [50, 5] when H_0 is true and Nominal Significance is $\alpha = .05$

$\sigma_1^2:\sigma_2^2$	PBS	PBP	PBB	PBM	UBS	UBP	UBB	UBM	TH	t	t'	RR
1:10	.411	.394	.403	.407	.151	.159	.162	.167		.395	.057	.129
1:5	.319	.319	.317	.327	.149	.155	.155	.153		.301	.057	.131
1:1	.058	.059	.059	.061	.134	.141	.143	.147		.049	.057	.130
5:1	.000	.000	.000	.001	.099	.101	.105	.107		.000	.059	.082
10:1	.001	.000	.000	.000	.079	.083	.078	.085		.000	.055	.061

Table 12

Actual Probability of Type I Error for Sample sizes [50, 50] when H_0 is true and Nominal Significance is $\alpha = .05$

$\sigma_1^2:\sigma_2^2$	PBS	PBP	PBB	PBM	UBS	UBP	UBB	UBM	TH	t	t'	RR
1:10	.055	.055	.057	.061	.057	.055	.055	.057		.051	.049	.054
1:5	.051	.051	.055	.063	.051	.055	.052	.055		.045	.044	.050
1:1	.057	.057	.059	.063	.056	.053	.058	.057		.052	.052	.051

Table 13

Actual Probability of Type I Error for Sample sizes [100, 20] when H_0 is true and Nominal Significance is $\alpha = .05$

$\sigma_1^2:\sigma_2^2$	PBS	PBP	PBB	PBM	UBS	UBP	UBB	UBM	TH	t	t'	RR
1:5	.234	.233	.231	.242	.067	.067	.069	.066	.220	.226	.046	.058
1:1	.050	.053	.055	.056	.067	.063	.063	.067		.048	.052	.056

Table 14

Actual Probability of Type I Error for Sample sizes [200, 20] when H_0 is true and Nominal Significance is $\alpha = .05$

$\sigma_1^2:\sigma_2^2$	PBS	PBP	PBB	PBM	UBS	UBP	UBB	UBM	TH	t	t'	RR
1:10	.395	.389	.382	.394	.062	.063	.064	.070		.389	.045	.063
1:5	.280	.275	.275	.282	.069	.067	.071	.075		.276	.047	.062
1:1	.054	.055	.054	.055	.077	.074	.076	.079		.052	.058	.074
5:1	.001	.000	.001	.000	.064	.065	.065	.068		.000	.054	.056
10:1	.000	.000	.000	.000	.051	.054	.055	.062		.000	.046	.059

Table 15

Actual Probability of Type I Error for Sample sizes [200, 200] when H_0 is true and Nominal Significance is $\alpha = .05$

$\sigma_1^2:\sigma_2^2$	PBS	PBP	PBB	PBM	UBS	UBP	UBB	UBM	TH	t	t'	RR
1:10	.051	.050	.052	.057	.052	.051	.050	.054		.051	.050	.053
1:5	.052	.053	.051	.054	.051	.053	.051	.059		.052	.052	.051
1:1	.053	.051	.050	.056	.049	.050	.050	.057		.050	.050	.054

Appendix C

The complete output for the initial analysis (power analysis with unadjusted α_n) in Phase Two simulation results are shown in Tables 16 through 24. The table column headings mimic those in Kulkarni (1993), and explanations for them can be found in Appendix B.

Table 16

Actual Power for Sample sizes [2, 2] when Nominal Significance is $\alpha = .05$

$\sigma_1^2:\sigma_2^2$	PBS	PBP	PBB	PBM	UBS	UBP	UBB	UBM	TH	t	t'	RR
1:10	.7620	.7645	.7595	.7860	.7600	.8140	.8025	.8135		.3030	.1515	.8140
1:5	.7410	.7450	.7240	.7700	.7425	.7835	.7760	.9865		.2600	.1290	.7835
1:1	.9975	.9980	.9920	.9980	.9980	.9990	.9850	.9990		.6780	.4040	.9990

Table 17

Actual Power for Sample sizes [5, 3] when Nominal Significance is $\alpha = .05$

$\sigma_1^2:\sigma_2^2$	PBS	PBP	PBB	PBM	UBS	UBP	UBB	UBM	TH	t	t'	RR
1:10	.4865	.4850	.4885	.4905	.4050	.4210	.4170	.4510		.3155	.1375	.3275
1:1	.8910	.8935	.8865	.8995	.8955	.8990	.8925	.9090		.7120	.6085	.7610

Table 18

Actual Power for Sample sizes [5, 5] when Nominal Significance is $\alpha = .05$

$\sigma_1^2:\sigma_2^2$	PBS	PBP	PBB	PBM	UBS	UBP	UBB	UBM	TH	t	t'	RR
1:10	.3550	.3520	.3555	.3615	.3560	.3625	.3505	.3735		.2120	.1710	.2775
1:5	.3270	.3220	.3215	.3285	.3240	.3270	.3230	.3380		.1865	.1655	.2510
1:1	.8460	.8440	.8415	.8530	.8425	.8420	.8420	.8490		.7040	.6700	.7985

Table 19

Actual Power for Sample sizes [20, 2] when Nominal Significance is $\alpha = .05$

$\sigma_1^2:\sigma_2^2$	PBS	PBP	PBB	PBM	UBS	UBP	UBB	UBM	TH	t	t'	RR
1:10	.6325	.6245	.6215	.6280	.4995	.5605	.5630	.5645		.5875	.1325	.5465
1:5	.5100	.5135	.5100	.5215	.4635	.5225	.5210	.5220		.4590	.1565	.5054
1:1	.7600	.7695	.7545	.7675	.8690	.8970	.8965	.8970		.6985	.3715	.8720
5:1	.0495	.0540	.0595	.0610	.5450	.5525	.5505	.5535		.0265	.3095	.4105
10:1	.0320	.0360	.0410	.0375	.6305	.6300	.6300	.6385		.0140	.3850	.4380

Table 20

Actual Power for Sample sizes [50, 5] when Nominal Significance is $\alpha = .05$

$\sigma_1^2:\sigma_2^2$	PBS	PBP	PBB	PBM	UBS	UBP	UBB	UBM	TH	t	t'	RR
1:10	.5575	.5475	.5450	.5565	.2600	.2665	.2710	.2760		.5420	.1080	.2185
1:5	.4575	.4605	.4595	.4620	.2565	.2710	.2670	.2715		.4425	.1075	.2385
1:1	.7520	.7500	.7430	.7575	.7845	.7855	.7840	.7940		.7190	.5525	.7660
5:1	.0130	.0275	.0335	.0325	.4100	.4160	.4160	.4245		.0230	.2900	.3615
10:1	.0110	.0125	.0135	.0155	.5205	.5210	.5155	.5275		.0080	.4250	.4145

Table 21

Actual Power for Sample sizes [50, 50] when Nominal Significance is $\alpha = .05$

$\sigma_1^2:\sigma_2^2$	PBS	PBP	PBB	PBM	UBS	UBP	UBB	UBM	TH	t	t'	RR
1:10	.1890	.1870	.1915	.1995	.1860	.1860	.1905	.2000		.1795	.1745	.1710
1:5	.1865	.1895	.1885	.1960	.1865	.1870	.1840	.1980		.1750	.1720	.1680
1:1	.7210	.7205	.7205	.7275	.7210	.7195	.7220	.7265		.7095	.7090	.6915

Table 22

Actual Power for Sample sizes [100, 20] when Nominal Significance is $\alpha = .05$

$\sigma_1^2:\sigma_2^2$	PBS	PBP	PBB	PBM	UBS	UBP	UBB	UBM	TH	t	t'	RR
1:5	.3890	.3895	.3885	.3995	.1630	.1570	.1615	.1680		.3805	.1290	.1605
1:1	.7350	.7355	.7295	.7370	.7360	.7305	.7280	.7420		.7230	.7000	.7175

Table 23

Actual Power for Sample sizes [200, 20] when Nominal Significance is $\alpha = .05$

$\sigma_1^2:\sigma_2^2$	PBS	PBP	PBB	PBM	UBS	UBP	UBB	UBM	Th	t	t'	RR
1:10	.5445	.5405	.5360	.5440	.1615	.1645	.1620	.1715		.5370	.1235	.1445
1:5	.4500	.4490	.4475	.4555	.1615	.1595	.1615	.1705		.4460	.1295	.1510
1:1	.7155	.7155	.7080	.7190	.7205	.7195	.7110	.7315		.7085	.6630	.7010
5:1	.0230	.0205	.0235	.0275	.3325	.3330	.3335	.3435		.0215	.3015	.3105
10:1	.0080	.0085	.0090	.0110	.4800	.4755	.4700	.4880		.0065	.4570	.4250

Table 24

Actual Power for Sample sizes [200, 200] when Nominal Significance is $\alpha = .05$

$\sigma_1^2:\sigma_2^2$	PBS	PBP	PBB	PBM	UBS	UBP	UBB	UBM	Th	t	t'	RR
1:10	.1965	.1940	.1950	.1955	.1970	.1965	.1920	.2030		.1965	.1965	.1680
1:5	.1760	.1805	.1790	.1900	.1740	.1785	.1770	.1825		.1745	.1740	.1500
1:1	.7055	.6960	.7010	.7115	.7050	.7065	.6960	.7060		.7055	.7055	.6850

Appendix D

The complete output for the second analysis (power analysis with adjusted α_n) in Phase Two simulation results are shown in Tables 25 through 33. The table column headings mimic those in Kulkarni (1993), and explanations for them can be found in Appendix B.

Table 25

Actual Power for Sample sizes [2, 2] when Actual Significance is $\alpha = .05$ for Student's t-test

$\sigma_1^2:\sigma_2^2$	PBS	PBP	PBB	PBM	UBS	UBP	UBB	UBM	Th	t	t'	RR
1:10	.7445	.7485	.7430	.7725	.7445	.8140	.8065	.8115		.2595	.1215	.8140
1:5	.7240	.7245	.7205	.7415	.7205	.7805	.7705	.7755		.2060	.0915	.7805
1:1	.9975	.9980	.9920	.9980	.9980	.9990	.9850	.9990		.6780	.4040	.9990

Table 26

Actual Power for Sample sizes [5, 3] when Actual Significance is $\alpha = .05$ for Student's t-test

$\sigma_1^2:\sigma_2^2$	PBS	PBP	PBB	PBM	UBS	UBP	UBB	UBM	TH	t	t'	RR
1:10	.3580	.3615	.3585	.3685	.2975	.3505	.3490	.3545		.1555	.0570	.3275
1:5	.8910	.8935	.8865	.8995	.8955	.8990	.8925	.9090		.7120	.6085	.7610

Table 27

Actual Power for Sample sizes [5, 5] when Actual Significance is $\alpha = .05$ for Student's t-test

$\sigma_1^2:\sigma_2^2$	PBS	PBP	PBB	PBM	UBS	UBP	UBB	UBM	Th	t	t'	RR
1:10	.3235	.3215	.3190	.3300	.3270	.3365	.3345	.3440		.1845	.1470	.2775
1:5	.3270	.3220	.3215	.3285	.3240	.3270	.3230	.3380		.1865	.1655	.2510
1:1	.8460	.8440	.8415	.8530	.8425	.8420	.8420	.8490		.7040	.6700	.7985

Table 28

Actual Power for Sample sizes [20, 2] when Actual Significance is $\alpha = .05$ for Student's t-test

$\sigma_1^2:\sigma_2^2$	PBS	PBP	PBB	PBM	UBS	UBP	UBB	UBM	Th	t	t'	RR
1:10	not run											
1:5	not run											
1:1	.7600	.7695	.7545	.7675	.8690	.8970	.8965	.8970		.6985	.3715	.8720
5:1	.0865	.0985	.0985	.0990	.5785	.5825	.5805	.5895		.0510	.3625	.3950
10:1	.0860	.0910	.0990	.0940	.7120	.7140	.7090	.7140		.0570	.5210	.4345

Table 29

Actual Power for Sample sizes [50, 5] when Actual Significance is $\alpha = .05$ for Student's t-test

$\sigma_1^2:\sigma_2^2$	PBS	PBP	PBB	PBM	UBS	UBP	UBB	UBM	TH	t	t'	RR
1:10	not run											
1:5	not run											
1:1	.7520	.7500	.7430	.7575	.7845	.7855	.7840	.7940		.7190	.5525	.7660
5:1	.0670	.0685	.0700	.0795	.4990	.4995	.4985	.5075		.0565	.3975	.3530
10:1	.5125	.5155	.5155	.5235	.8820	.8810	.8795	.8810		.4955	.8655	.4255

Table 30

Actual Power for Sample sizes [50, 50] when Actual Significance is $\alpha = .05$ for Student's t-test

$\sigma_1^2:\sigma_2^2$	PBS	PBP	PBB	PBM	UBS	UBP	UBB	UBM	Th	t	t'	RR
1:10	.1890	.1870	.1915	.1995	.1860	.1860	.1905	.2000		.1795	.1745	.1710
1:5	.1865	.1895	.1885	.1960	.1865	.1870	.1840	.1980		.1750	.1720	.1680
1:1	.7210	.7205	.7205	.7275	.7210	.7195	.7220	.7265		.7095	.7090	.6915

Table 31

Actual Power for Sample sizes [100, 20] when Actual Significance is $\alpha = .05$ for Student's t-test

$\sigma_1^2:\sigma_2^2$	PBS	PBP	PBB	PBM	UBS	UBP	UBB	UBM	TH	t	t'	RR
1:5	not run											
1:1	.7350	.7355	.7295	.7370	.7360	.7305	.7280	.7420		.7230	.7000	.7175

Table 32

Actual Power for Sample sizes [200, 20] when Actual Significance is $\alpha = .05$ for Student's t-test

$\sigma_1^2:\sigma_2^2$	PBS	PBP	PBB	PBM	UBS	UBP	UBB	UBM	TH	t	t'	RR
1:10	not run											
1:5	not run											
1:1	.7155	.7155	.7080	.7190	.7205	.7195	.7110	.7315		.7085	.6630	.7010
5:1	.3475	.3430	.3395	.3615	.6900	.6870	.6830	.6960		.3440	.6785	.3170
10:1	.4740	.4765	.4815	.4815	.8445	.8415	.8455	.8470		.4670	.8420	.4165

Table 33

Actual Power for Sample sizes [200, 200] when Actual Significance is $\alpha = .05$ for Student's t-test

$\sigma_1^2:\sigma_2^2$	PBS	PBP	PBB	PBM	UBS	UBP	UBB	UBM	TH	t	t'	RR
1:10	.1965	.1940	.1950	.1955	.1970	.1965	.1920	.2030		.1965	.1965	.1680
1:5	.1760	.1805	.1790	.1900	.1740	.1785	.1770	.1825		.1745	.1740	.1500
1:1	.7055	.6960	.7010	.7115	.7050	.7065	.6960	.7060		.7055	.7055	.6850

Appendix E

The complete output for the third analysis (power with effect size maintained) in Phase Two simulation results are shown in Table 34. The table column headings mimic those in Kulkarni (1993), and explanations for them can be found in Appendix B.

Table 34

Actual Power for Various Sample sizes when Effect Size is Maintained

n_1	n_2	$\sigma_1^2:\sigma_2^2$	PBS	PBP	PBB	PBM	UBS	UBP	UBB	UBM	t	t'	RR
20	2	5:1	.3285	.3260	.3310	.3450	.8655	.8670	.8710	.8705	.2370	.5475	.7425
20	2	10:1	.8195	.8125	.7860	.8240	.9995	.9995	.9990	.9995	.7390	.9075	.9830
50	50	1:5	.7100	.7085	.7045	.7180	.7090	.7065	.7050	.7120	.6920	.6885	.6390

Appendix F

When Student's t-test is found inappropriate, a nonparametric technique often used to test the difference between the medians of two independent samples is the Robust Rank Order (RRO) test. The Robust Rank Order test requires only that the data be of ordinal measure and is based on the rank order of the observed data. The Robust Rank Order test tests for equality of medians rather than means, but it has been successfully used as a comparison to Student's t-test in studies of Type I error rates when the populations are known to be normally distributed (Kulkarni, 1993). The comparison is effective in such cases because the normal distribution ensures that the mean and median are equal. Further, studies have shown that this test controls α_a well when sample sizes are large, whether or not they are equal (Kulkarni, 1993). In particular, the RRO test is insensitive to heterogeneity of variance even when sample sizes are not equal.

The nonparametric Robust Rank Order test works as follows: The data from the two samples, Y_1 and Y_2, are combined and listed in ascending order. For each score from sample Y_1, the number of scores from Y_2 which have a lower rank are counted and denoted by $U(Y_2\ Y_{i1})$ where $i = 1, 2, ..., n_1$. The same is done for each score from sample Y_2, and the results are denoted by $U(Y_1 Y_{j2})$ where $j = 1, 2, ..., n_2$. The means for $U(Y_2\ Y_{i1})$ and $U(Y_1\ Y_{j2})$ are then calculated and denoted by $U(Y_2\ Y_1)$ and $U(Y_1\ Y_2)$, respectively. From these, two indices of variability, V_{y1} and V_{y2}, are calculated.

$$V_{y1} = \sum_{i=1}^{n_1}[U(Y_2Y_{i1}) - U(Y_2Y_1)]^2 \qquad V_{y2} = \sum_{i=1}^{n_2}[U(Y_1Y_{j2}) - U(Y_1Y_2)]^2$$

The test statistic is then computed as follows:

$$U' = \frac{n_1U(Y_2Y_1) - n_2U(Y_1Y_2)}{2[V_{y1} + V_{y2} + U(Y_1Y_2)U(Y_2Y_1)]^{1/2}}$$

Finally, reference is made to a standard table to see if the value of U' is significant, suggesting that the population medians may not be equal.

Appendix G

The following is the actual program implementation from DATASIM used for generating all of the data sets used in Phase One of this study. Commands are printed in capital letters.

COPY 666 BUF0 < 1>, exact on

DESIGN TWOGROUP, NOBS n_1 n_2,MU 0,SIGMA σ_1 σ_2,DECI 10

TIME, CLOSE SCREEN, OPEN (temporary-datafile-name)

SIMU 2000 -999,WINDOW-----. ##########, \

COPY BUF0 (results-datafile-name) <2>, COPY DATA (results-datafile-name) <2>,\

TWOT 01, HETT 01;

CLOSE (temporary-datafile-name), OPEN SCREEN, TIME

The following is the actual program implementation from DATASIM used for generating all of the data sets used in Phase Two of this steady. Commands are printed in capital letters.

COPY 666 BUF0 < 1>, exact on

DESIGN TWOGROUP, NOBS n_1 n_2, MU 0 μ_2, SIGMA σ_1 σ_2,DECI 10

TIME, CLOSE SCREEN, OPEN (temporary-datafile-name)

SIMU 2000 -999,WINDOW-----. ##########,\

COPY BUF0 (results-datafile-name) < 2 >, COPY DATA (results-datafile-name) < 2 > ,\

TWOT 01, HETT 01;

CLOSE (temporary-datafile-name), OPEN SCREEN, TIME

Vita

Laura L. Lansing was born and raised in Champaign, Illinois, the daughter of Dr. Kenneth M. and Alice K. Lansing. She received her Bachelor of Arts degrees from Rockford, Rockford, Illinois in Philosophy and Mathematics in 1982 and 1986, respectively. She received a Master of Science in Mathematics with a major in Operations Research from The College of William and Mary College in Virginia, Williamsburg, Virginia in 1989. She was awarded first prize in a national student paper contest at the 1992 ORSA Conference, San Francisco, California. She received a Master of Science in Industrial Engineering with a major in Operations Research from Lehigh University, Bethlehem, Pennsylvania in 1994. She has been employed as a teaching assistant during her tenure as a graduate student. Laura is currently pursuing a PhD. in Psychology from Lehigh University.

END OF TITLE